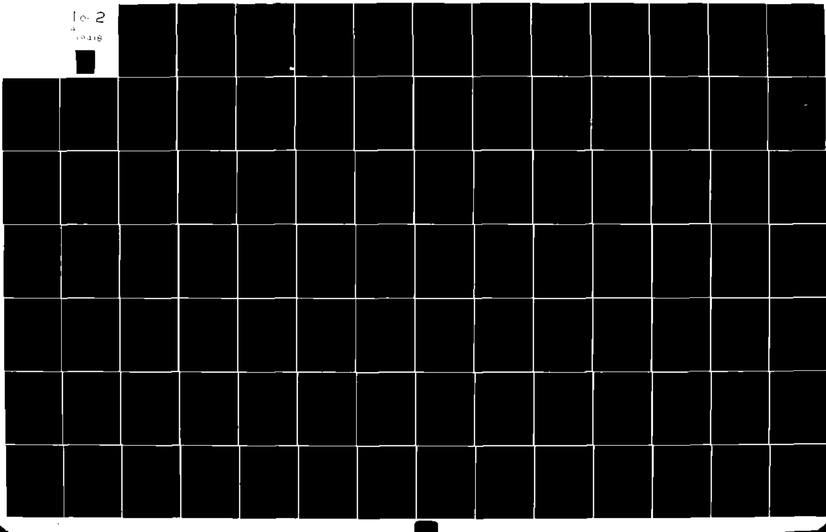
AD-A119 418 STANFORD UNIV CA DEPT OF COMPUTER SCIENCE F/G 12/1
WAVE PROPAGATION AND STABILITY FOR FINITE DIFFERENCE SCHEMES.(U)
MAY 82 L N TREFETHEN N00014-75-C-1132
UNCLASSIFIED STAN-CS-82-905 NL

1 of 2
AD
A119418



23

April 1982

Report. No. STAN-CS-82-905

Wave Propagation and Stability for Finite Difference Schemes

by

Lloyd N. Trefethen

Department of Computer Science

Stanford University
Stanford, CA 94305

A

WAVE PROPAGATION AND STABILITY FOR FINITE DIFFERENCE SCHEMES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

N00014-75-C-1132

A

By
Lloyd Nicholas Trefethen
May 1982

Abstract

This dissertation investigates the behavior of finite difference models of linear hyperbolic partial differential equations. Whereas a hyperbolic equation is nondispersive and nondissipative, difference models are invariably dispersive, and often dissipative too. We set about analyzing them by means of existing techniques from the theory of dispersive wave propagation, making extensive use in particular of the concept of *group velocity,* the velocity at which energy propagates.

The first three chapters present a general analysis of wave propagation in difference models. We describe systematically the effects of dispersion on numerical errors, for both smooth and parasitic waves. The reflection and transmission of waves at boundaries and interfaces are then studied at length. The key point for this is a distinction introduced here between *leftgoing* and *rightgoing* signals, which is based not on the characteristics of the original equation, but on the group velocities of the numerical model.

The last three chapters examine *stability* for finite difference models of *initial boundary value problems.* We show that the abstract stability criterion of Gustafsson, Kreiss, and Sundström (GKS) is equivalent to the condition that the boundary permit no rightgoing signals in the absence of leftgoing ones. Wave propagation arguments yield a proof that for the typical instability of "strictly rightgoing" type, one has unstable growth in the ℓ_2 norm, not just in the complicated GKS norm. We prove that this growth is at least proportional to the number of time steps n for models driven by boundary data, and to $\sqrt{n}$ for models driven by initial data.

We show further that most GKS-unstable boundaries exhibit *infinite reflection coefficients,* which gives an alternative explanation of instability with respect to initial data. We conjecture that when an infinite reflection coefficient is present, the unstable growth rate increases from $\sqrt{n}$ to n.

Throughout the dissertation, wave propagation ideas are also applied to various more specialized stability problems. We identify new classes of unstable formulas, including some in two space dimensions; derive new results relating stability to dissipativity; give new estimates on unstable growth for problems with two boundaries or interfaces; examine borderline cases that are GKS-unstable but ℓ_2-stable or nearly so; and present an explanation based on dispersion for known results on instability in L_p norms.

Acknowledgments

I am indebted to many people for help and advice given in the course of this research. They include David Brown, Russel Caflisch, Ray Chin, Jonathan Goodman, Bertil Gustafsson, Joseph Keller, Heinz Kreiss, John Strikwerda, Bob Warming, and Helen Yee. Financial support has been provided by a National Science Foundation Graduate Fellowship and a Hertz Foundation Fellowship, and also by Office of Naval Research Contract N00014-75-C-1132, National Science Foundation Grant MCS77-02082, and NASA-Ames University Consortium Interchange NCA2-OR745-004.

Technologically, writing the dissertation was made possible By Donald Knuth's superb mathematical typesetting system, TEX. Numerical computations were performed at the Stanford Linear Accelerator Center of the U.S. Dept. of Energy. Symbolic calculations of dispersion relations were checked by means of the MACSYMA system supported by the Mathlab Group, Laboratory for Computer Science, M.I.T.

I have been fortunate to carry out this work amidst a group of "numerical pde" people with closely similar interests. I gratefully thank my adviser, Joseph Oliger, for bringing together this group and for guiding our work with patience and good nature. The other students involved have been William Coughran, William Gropp, and especially important to me, Randall LeVeque and Marsha Berger, who have been invaluable friends and research companions. Other members of the Stanford Numerical Analysis Group who have helped make my experience pleasurable and productive have been Petter Bjørstad, Eric Grosse, and Robert Schreiber, who also made many valuable additions as a member of my reading committee. I will remember with pleasure my years at Stanford with all of these people.

I would like to express special additional thanks to two people. One is Gerald Hedstrom, of the Lawrence Livermore Laboratory, who has taken an extraordinarily active interest in my work from the start, pointing out several errors and bringing innumerable new ideas and references to my attention. It is hard to see how I could have written a thesis involving dispersive wave propagation without the benefit of Hedstrom's experience.

Secondly, the credit for making my graduate student years successful in a larger sense must go to Gene Golub, whose involvement and generosity have been unflagging. The Stanford Numerical Analysis group owes its intimacy and its energy principally to Gene, and to his active faith is the importance of numerical analysis.

The thesis is dedicated to Carolyn Gramlich, with whom I looked at stars and planets at the end of each evening of coffee and group velocity.

Table of Contents

List of Illustrations

List of Tables

Index of Notation

LF	Leap Frog (§1.1)	LW	Lax-Wendroff (§1.1)
CN	Crank-Nicolson (§1.1)	LF4	4th-order Leap Frog (§1.1)
BE	Backwards Euler (§1.1)	LFd	Dissipative Leap Frog (§1.1)
LF^2	Leap Frog for $u_{tt} = u_{xx}$ (§3.2)	UW	Upwind (§A)
LxF	Lax-Friedrichs (§A)	BX	Box (§A)
S	Space Extrapolation (§3.2)	ST	Space-time Extrapolation (§3.2)
MOL	Method of Lines (§A)		

$\mathbb{Z}, \mathbb{R}, \mathbb{C}$	integers, real numbers, complex numbers
x, t	independent space, time variables (§1.1)
$u = u(x,t)$	dependent variable (§1.1)
$u_t = au_x$	scalar model equation (§1.1)
ξ, ω	wave number, frequency (§1.1)
h, k	mesh size in x, t (§1.1)
Q	difference model (§1.1, 2.1, 2.5)
λ	mesh ratio k/h (§1.1)
v_j^n	difference approximation to $u(jh, nk)$ (§1.1)
α, β	order of dispersion, dissipation (§1.1)
c, C	phase speed ω/ξ, group speed $d\omega/d\xi$ (§1.2)
W, A	width, amplitude of wave packet (§1.3, 1.4)
d	number of space dimensions (§1.6)
$\mathbf{x}, \boldsymbol{\xi}$	position, wave number vectors (§1.6)
$\mathbf{c}, \mathbf{C}$	phase velocity, group velocity vectors (§1.6)
K, Z	shift operators $Kv_j^n = v_{j+1}^n$, $Zv_j^n = v_j^{n+1}$ (§2.1)
κ, z	amplification factors $\kappa = e^{-i\xi h}$, $z = e^{i\omega k}$ (§2.1)
δ	degree of defectiveness (§2.1)
ℓ, r, s	stencil parameters (§2.1, 2.5)
$P(K, Z)$	bivariate polynomial representation of Q (§2.1, 2.5)
$P_\kappa(Z), P_z(K)$	localised univariate polynomials $P(\kappa, Z), P(K, z)$ (§2.1, 2.5)
$\tilde{C}$	translation speed for signal with $\lvert\kappa\rvert \neq 1$ (§2.3)
ν_ℓ, ν_r	multiplicity of left-, rightgoing signals (§2.3)

κ_ℓ, κ_r	leftgoing, rightgoing amplification factors (§2.4)
$\rho(Z), \sigma(Z)$	polynomials defining 3-pt. linear multistep formula (§2.4)
N	dimension of vector system (§2.5)
$\mathbf{u}_t = A\mathbf{u}_x$	vector model equation (§2.5)
n_ℓ, n_r	no. of leftgoing, rightgoing signals (§2.5)
$\boldsymbol{\psi}$	vector wave (§2.5)
$\bar{Q}$	difference model with boundary (§3.1, 3.6, 4.2)
A, B	reflection, transmission coefficients (§3.1)
$A(z)$	reflection coefficient function (§3.1)
$\kappa_i, \kappa_r, \kappa_t$	incident, reflected, transmitted amplification factors (§3.1)
Φ	energy flux (§3.3)
ω_c	cutoff frequency (§3.4)
$\{S_{j\sigma}\}$	boundary conditions (§3.6)
$D^{[\ell]}(z), D^{[r]}(z)$	reflection matrices (§3.6)
S	solution operator (homog. boundary data) (§1.4, 4.2, B)
$S_{bc}^{(n)}$	solution operator (homog. initial data) (§4.2)
$\boldsymbol{\phi}$	eigensolution or generalised eigensolution (§4.2, 4.3)
f, g, F	initial, boundary, forcing data (§4.2)
$\boldsymbol{\kappa}$	spatial amplification factor vector (§4.5)

0. INTRODUCTION

0.1 Purpose

Many problems of physics and engineering take the form of *hyperbolic systems of partial differential equations* [Co62]. Some examples of fields in which these equations are important are fluid mechanics (weather prediction, aircraft and turbine design, oceanography...), geophysics (earth modeling, petroleum prospecting...), magnetohydrodynamics, elasticity, and acoustics. In most instances there is no hope of obtaining analytical solutions, and one must resort to numerical approximations. Of these the most important are the *finite difference models*, based on the idea of approximating partial derivatives by discrete differences.

An irony of the finite difference process, as is well known, is that the detailed behavior of finite difference formulas is generally a good deal more complicated than that of the differential equations they model. For the most part this is not a problem, because the nonphysical details are unimportant so long as the numerical solution converges to the correct physical result when the grid is refined. This convergence will normally take place provided that the difference model is *consistent* and *stable* [Ri67,Gu75]. Therefore the analysis of the behavior of difference models traditionally reduces to estimating truncation errors by Taylor expansions, in order to determine consistency and asymptotic accuracy, and to some kind of investigation of stability. Of these the stability analysis is the much more difficult task.

To check for stability in the case of linear problems with smooth coefficients and no boundaries, it is essentially enough to make sure that the difference formula admits no exponentially growing Fourier modes [Ri67,Th69]. But for problems with boundaries, as are almost always present in practice, the question becomes more difficult. One can still push through an analysis based on an extended notion of "growing modes," but it is not straightforward. A general theory of this kind was developed by Kreiss and colleagues a decade ago and was reported in an important

paper of Gustafsson, Kreiss, and Sundström—henceforth "GKS"—in 1972 [Gu72]. (See also [Co80,Gu75,Kr71,Mi81].) This theory is powerful, but mathematically and conceptually difficult. The proofs involved are obscure enough that it is fair to say that most people apply the GKS results without understanding them.

This dissertation develops the view that a finite difference model is not just a mathematical corruption of an ideal problem, but a physical medium of a different kind with analysable properties of its own. Finite difference models do not exhibit the characteristic features of hyperbolicity, such as finite speed of propagation. Instead, they act as *dispersive media*, a subject about which a great deal is known [Br60,Li78,Wh74]. Wave propagation in such media is characterised by *dispersion* of different frequencies and by energy propagation at a frequency-dependent speed called the *group velocity*. These effects depend on the interference of distinct frequency components, and therefore represent a step beyond the superposition of individual Fourier modes. Our contention is that dispersive wave propagation phenomena are the essential feature underlying much of the more subtle behavior of difference models. In particular, the GKS stability theory has a simple physical explanation in terms of group velocity.

Our interpretation of the main GKS result runs roughly as follows. Let a difference model for an initial boundary value problem be applied with homogeneous boundary data. To be stable, the model must admit no solutions that grow exponentially in the number of time steps (a result first exploited by Godunov and Ryabenkii [Ri67]). But in addition, it must admit no solutions consisting of a collection of waves radiating from the boundary into the interior. Such waves might be *physical* (i.e. smooth, close to waves admitted by the differential equation), or *parasitic* (not smooth), but this distinction does not appear in the analysis. For a wave to propagate "into the interior" means, in the case of a boundary at the left of a region, for it to have a positive group velocity.

The analysis also makes no explicit distinction between dissipative and nondissipative difference formulas. Dissipativity, however, guarantees a priori that most wavelike modes cannot occur, and this limits the range of potential radiating solutions that must be investigated in checking for stability.

Thus we show that instability for initial boundary value problems is a kind of resonance phenomenon, in which some energy-radiating solution can oscillate continually at the boundary without being continually forced by inhomogeneous boundary data or by signals hitting the boundary from the interior. The question arises as to the extent to which such resonance will be excited by rounding *errors, truncation errors, or* other data. Regarding stimulation by boundary data, we conclude that the resonance will in general always be excited. But for stimulated resonance by initial or field data, the matter of *reflection coefficients* becomes important. Indeed one purpose of this dissertation is to demonstrate how closely stability for initial boundary value problems is tied, both formally and physically, to reflection phenomena. We show that the "standard" GKS-instability is characterized by infinite reflection coefficients, leading to great sensitivity of the solution to energy hitting the boundary, but that there are realistic borderline cases with finite or zero reflection coefficients, and in these the instability is not so easily excited.

* * *

Several difficulties have inhibited the theoretical and practical application of the GKS theory. One, as mentioned above, is that the mathematics involved is complicated and not clearly motivated. We hope that the wave propagation point of view can remove some of this mystery. A second is that the GKS stability definition is complicated and unnatural—it gives estimates in a norm that one would not normally be interested in. We will show that the group velocity analysis allows one to derive estimates for most unstable cases in the simpler ℓ_2 norm. How best to measure stability for models of initial boundary value problems is however a complicated question, to which there is no universal answer, and we will attempt to shed light on it by a variety of examples and arguments. A third difficulty is that the algebraic process of testing for instability can be extremely difficult for nontrivial initial boundary value problem models [Co80]. Fundamentally our ideas do not help with this problem at all. There is probably not much to be done about this in general, we believe, as the algebra reflects a physical behavior that is truly complex. However, results will be given that shortcut the analysis for special classes of problems.

The "wave propagation" approach to stability might be contrasted with the more standard "semigroup" point of view. The latter considers difference models as time-evolution operators, and characteristically investigates what "growth" can take place from one time step to the next. The former views space and time more equally, and investigates what qualitative changes occur between time steps—which may indeed cause growth, but indirectly.

The wave propagation view is not always easy to shape into mathematical proofs. As a general rule, one can prove instability and determine a lower bound for its magnitude by studying unstable waves with behavior regular enough for asymptotic analysis. This is what we have done for the ℓ_2 results mentioned above. Proving stability, on the other hand, or establishing upper bounds for unstable growth rates, takes a greater effort, because it requires consideration of arbitrary signals with no regular behavior.

As dispersive media with a periodic structure, finite difference models have a great deal in common with solid crystals (and also with certain other periodic physical systems, such as regular electric networks). Accordingly, the general features of wave propagation that we will discuss have close analogs in the solid state physics literature [Bo54,Br53,Ma69,So64]. However, the analogy is least close in the area of stability, which corresponds approximately to energy conservation for physical systems. For crystals, energy conservation is one of the postulates from which local solution behavior may be derived, while in our context, it is the local behavior that is given and the stability that is under question. (See, however, Part III of [Bo54].)

* * *

Three main themes will occupy us throughout the dissertation:

(A) group velocity and parasitic waves... leftgoing and rightgoing solutions;

(B) reflection and transmission at boundaries and interfaces;

(C) stability.

Our first three chapters are devoted to an exposition of the phenomena (A) and (B) and their relationship. Some of our results are old, but many are new, and this is the most systematic presentation of such material that has appeared to date. The last three chapters are concerned with stability theory (C) for initial boundary value problems. They present our analysis of the GKS theory as an outgrowth of (A) and (B). This leads to new results of various kinds. For a detailed outline see §0.3, below.

The general purpose of this dissertation is to shed new light on the existing theory of finite difference models, and to extend the theory where possible. However, we suspect that most fruitful applications of the wave propagation point of view potentially lie in more novel and difficult areas that are only touched on here, such as problems with variable coefficients, nonlinear problems, problems with characteristic boundaries, and multidimensional problems with irregular boundaries. If our belief is valid that the essential features of discretization for hyperbolic problems are those of dispersive wave propagation, then further work on these lines ought to point the way to new and hitherto unrecognized phenomena.

0.2 History

Regarding the application of ideas of dispersive wave theory to the theory of difference models, I am aware of two important sets of predecessors. The first are G. Hedstrom and R. Chin, who in a variety of papers have applied wave theory arguments to analyse many aspects of solution behavior and (Cauchy) stability [He65,He66,He68, He75,Ch75,Ch78,Ch79,Ch83]. Making extensive use of saddle-point estimates, these papers study stability in the maximum norm (see §1.4), analysis by modified equations (see §§1.1, 1.2), and solution behavior near discontinuities. The second are R. Vichnevetsky and his colleagues, who for a particular semi-discrete model of $u_t = u_x$ (usually), analyse wave propagation for both smooth and parasitic waves [Vi75,etc.]. Vichnevetsky's papers do not perform explicit saddle-point analysis, and as a result they do not obtain the kind of precise estimates derived by Hedstrom and Chin. However, his interest in parasitic waves and in behavior at boundaries makes these papers the most direct precursor to this dissertation. Vichnevetsky's work will be summarized shortly in a book with J. Bowles [Vi82].

Besides these, there are undoubtedly a large number of group velocity calculations for difference models in the literature, most of which I am probably unaware of. To the authors of these I apologize in advance. Three references that I do know, from geophysics, are the reports of Alfold, et al. [Al74], Bamberger, et al. [Ba80], and Martineau-Nicoletis [Ma81]. These works are mainly concerned with smooth waves rather than parasites; the first treats the acoustic (standard) wave equation, and the other two the elastic wave equation (pressure and shear).

Similarly, there are no doubt a number of papers that compute numerical reflection and transmission *coefficients for boundaries or interfaces, as done here in* §3 and thereafter. I am aware of such calculations by Martineau-Nicoletis [Ma81], D. Brown [Br79,Cl79], and Vichnevetsky [Vi81b]. Only Vichnevetsky makes a connection with group velocity. The general description presented here of behavior at an interface in terms of left- and rightgoing waves admitted on either side appears to be new.

The stability theory for initial boundary value problems that is the main concern

here has a complicated history. The dissertation refers primarily to the paper of Gustafsson, Kreiss, and Sundström [Gu72], which seems to have dominated the field since its appearance in 1972. However, this emphasis does not do justice to many important contributions by G. Strang, S. Osher, and others. In particular, Osher's paper [Os69b] obtains a large part of the main GKS result by different means. Osher considers only models that satisfy a certain root-separation condition, which rules out many nondissipative difference formulas (those admitting a wave with group velocity 0); on the other hand, his result has the advantage of using the ℓ_2 norm rather than the more unwieldy GKS stability definition.

Here is a very brief survey of the history of stability theory for difference models of initial boundary value problems. The first contributions were made by Godunov and Ryabenkii in the early 1960's, who observed that a necessary condition for stability is that the spectrum of the time-evolution difference operator be contained in the unit disk in the limit as the mesh size becomes 0, and derived conditions for this to occur [Ri67]. This is the beginning of the use of *normal mode analysis* in stability theory for initial boundary value problems, which pervades the subsequent results. The Godunov-Ryabenkii condition is an analog for initial boundary value problems of the von Neumann condition for initial value problems, and like the von Neumann condition, it is necessary for stability but not sufficient. The next contributions were due to Strang and to Kreiss. Strang applied a factorization technique for Toeplitz matrices, related to the Wiener-Hopf method, to obtain necessary and sufficient stability conditions for a restricted set of difference approximations, namely those with purely homogeneous boundary conditions [St64,St66]. By different methods, Kreiss [Kr66] obtained a sufficient condition for stability of diagonalizable (essentially scalar) two-level explicit dissipative models. In [Os69a], Osher proved a similar result by an extension of Strang's approach, introducing general boundary conditions by means of a finite-rank correction to the Toeplitz operator for the interior difference scheme.

These papers left two main gaps in the available theory. First, they did not say much about nondissipative models. Second, they did not deal with nondiagonalizable models. In another paper published in 1969, Osher made some progress on the first problem, again by the Toeplitz factorization technique, obtaining a result that weakens dissipativity to a separation-of-roots condition [Os69b]. This was a quite general theorem along the lines of "the absence of eigensolutions and generalized eigensolutions ensures stability," which we will discuss in §4. Kreiss, on the other hand, derived a sufficient condition for stability of dissipative nondiagonalizable models in [Kr68], by making use of a Dunford integral to bound the powers of the discrete time-evolution operator.

It remained to derive a stability condition for general nondissipative models, and if possible, one that would be necessary as well as sufficient. The groundwork for this was work by Kreiss on matrix normal forms for initial boundary value problems for partial differential equations (not difference models), published in [Kr70]. These results led to necessary and sufficient conditions for well-posedness of hyperbolic partial differential equations in several space dimensions. By an extension of the same ideas, the paper of Gustafsson, Kreiss and Sundström [Gu72] finally proved a general necessary and sufficient stability theorem for (one-dimensional) difference models, dissipative or nondissipative, diagonalizable or nondiagonalizable.

Further additions to the stability theory since 1972 have mainly taken the form of embellishments of the GKS theory. Gustafsson in [Gu75] established connections between GKS-stability and convergence; the main problem here is working around the idiosyncrasies of the GKS stability definition so as to be able to treat nonzero initial data. Ciment [Ci71,Ci72], Burns [Bu78], Tadmor [Tad81], and Goldberg and Tadmor [Ta78,Go78,Go81] have proved additional results. GKS-like theorems have been obtained for method-of-lines schemes by Strikwerda [St78], and for parabolic problems by Varah [Va70,Va71] and Osher [Os72]. Most recently, attention has shifted to problems in several space dimensions [Co80,Mi81]; in particular, new results of Michelson's [Mi81] offer promise of a complete extension of the GKS theory to dissipative multidimensional models. In addition, there have been numerous papers that apply the GKS theory to study stability of particular difference formulas or classes of them, including [Ab79,Ab81,Be81,Br73,Co80,Go78b,Ol74,Ol76,Su74].

Virtually all of these results, both preceding and following [Gu72], can be given wave propagation interpretations. For example, several of them amount to statements *that spontaneous radiation from the boundary implies instability*, but with the radiation restricted to zero-frequency components that correctly mimic the differential *equation, instead of the more general posibility of parasitic waves radiating energy* according to the group velocity [Bu78,Kr68,Ta81]. None of them are presented in this way, but the *relevance to stability of* "energy propagating in the wrong direction" is mentioned in some of Kreiss's papers. In at least two places he performs a calculation

in which the parasitic solution of one difference formula is related to the smoot solution of another, whose speed of propagation *is then obvious by consistenc*; tl is a calculation of group velocity in disguise ([Br73] or [Kr73], §20; [Kr74] or [Kr7. §17). In an early paper with Lundqvist [Kr68b], Kreiss also defines the *concept* *strictly noncontractive* difference formulas in terms of a quantity that is group veloci without the name (see also [Ap68] and [Os69c]). *In fact, Thm. 4 of* [Kr68b] is v closely related to Thm. 4.2.3 here. However, it seems clear that the central posit of group velocity in stability theory *has not been seen before; to my knowledge*, t words "stability" and "group velocity" have not appeared together in the past.

0.3 Outline and summary of results

This dissertation is unfortunately quite lengthy, as the following detailed outline makes clear. To mitigate this problem somewhat, a general index is provided at the end. Readers wishing to go as quickly as posible to the stability theory for initial boundary value problems should proceed to Chapter 4 after reviewing Sections 1.1, 1.2, 1.5, 2.3, and 3.1. For a quick view of our main stability ideas, see Sections 4.1, 4.2, and 5.5. Published accounts corresponding roughly to Chapters 1 and 4 can be found in [Tr82] and [Tr83], respectively.

Chapter 1. We begin in §1 with a discussion of the behavior as dispersive media of finite difference models of the scalar equation $u_t = au_x$. Our model approximates $u(x,t) = u(jh, nk)$ by a quantity v_j^n, where h and k are the ***space step size*** and ***time step size***. In §1.1 we define the concepts of ***frequency*** ω, ***wave number*** ξ, and ***dispersion relations***, and relate these to ***consistency, accuracy***, and ***modified equations***. We illustrate these ideas by applying them to a number of well-known difference formulas, which continue to serve as examples throughout the dissertation. (These are summarized in Appendix A.) Section 1.2 defines ***phase speed*** $c(\xi,\omega)$ and ***group speed*** $C(\xi,\omega)$, and derives the latter by the ***method of stationary phase***. The effect of group velocity is illustrated by numerical experiments involving ***wave packets*** and ***wave fronts***. Thm. 1.2.1 points out that for a general nondissipative difference model, errors in C are greater than errors in c by a factor equal to the ***order of dispersion***. Section 1.3 shows the connection between group velocity and ***dispersion***, with further numerical illustrations. In §1.4 we apply these ideas to show that certain known results on ***L_p-instability*** of difference models for $p \neq 2$ can be explained quantitatively in terms of dispersion and dissipation. In §1.5 we examine parasitic waves, and show that they too are governed by a group velocity. More numerical illustrations are given. New concepts of *x-reversing* and *t-reversing* formulas are introduced and applied in Thm. 1.5.1, and Thm. 1.5.2 shows that most nondissipative formulas are x- or t-reversing. Section 1.6 briefly surveys wave propagation in ***multidimensional difference models***, where energy propagation is governed by a ***vector group velocity*** C and wave packets can be tracked by a process of ***numerical ray tracing***. Some of these ideas are new, but we do not develop them. (More details can be found in [Tr82].)

Chapter 2. Chapter 2 sets out to make the ideas of §1 more general and more rigorous. In §2.1 we define the general constant-coefficient scalar difference formula Q in terms of shift operators K and Z, and analyze what solutions it supports that are regular in x or t (Thms. 2.1.1,2.1.2). In addition to ξ and ω, we now begin to work with arbitrary complex *space* and *time variation factors* $\kappa = e^{-i\xi h}$ and $z = e^{i\omega k}$. The new concept of a *separable* formula is defined, and it is shown that for separable formulas, $C(\xi,\omega)$ factors into $C_1(\xi)C_2(\omega)$. Section 2.2 defines *Cauchy stability* and relates this to the *von Neumann condition* and a *root condition* (Thm. 2.2.1). It also defines *(x)-dissipativity* and relates this to the new concepts of *t-dissipativity* and *total dissipativity* (Thms. 2.2.2,2.2.3). Thm. 2.2.4 points out that if Q is x- or t-dissipative, it cannot be x- or t-reversing. In §2.3 we establish that the group velocity makes sense in a general way by proving that every wave admitted by any Cauchy stable formula, whether dissipative or nondissipative, has a group velocity (Thm. 2.3.1). Thm. 2.3.2 proves further that C is the limit of the translation speeds $\hat{C}$ of evanescent waves, and that the sign of C can be determined by a *perturbation test*. We also define the new concepts of *stationary*, *rightgoing* and *strictly rightgoing*, *leftgoing* and *strictly leftgoing* signals in terms of group velocity, and these are summarized in Table 2.1. Section 2.4 applies most of the results up to that point to the interesting case of ***three-point linear multistep formulas*** studied by Beam, Warming, and Yee [Be79,Be81]. New results are proved relating *A-stability* and *strong A-stability* of such formulas to their wave propagation behavior (Thm. 2.4.1) and t-dissipativity (Thm. 2.4.2). Finally, Section 2.5 shows that all of the results established for scalar models carry over directly to diagonalizable systems. In particular, Thm. 2.5.1 describes the general breakdown of time-regular vector solutions into leftgoing and rightgoing components.

In summary, Chapters 1 and 2 present the essentials of dispersive wave theory for finite difference models in the absence of boundaries, and document the importance of this theory by showing its many effects theoretically and with numerical demonstrations. The most original ideas here are those related to multidimensional problems (§1.6) and L_p-instability (§1.4). None of the results have much technical depth; perhaps the least trivial is the general justification of group velocity in Thms. 2.3.1 and 2.3.2.

Chapter 3. In §3 we begin to deal with boundaries and interfaces. Section 3.1 describes our general procedure for computing ***reflection and transmission coefficients*** for ***steady-state solutions*** of the form $v^n = z^n v^0$: first determine all ***leftgoing*** and ***rightgoing*** signals admitted away from the interface, as defined in §2, then match these by algebraic interface conditions. This procedure depends upon a numerical analog of the *Sommerfeld radiation condition*. Section 3.2 computes reflection and transmission formulas for a large number of examples involving both boundaries and interfaces, and verifies two of these with numerical experiments; the most complicated example involves an abrupt change between two arbitrary difference formulas, for which a van der Monde matrix comes into play. Section 3.3 considers ***energy conservation*** at interfaces, and §3.4 discusses ***cutoff frequencies*** and ***stop bands***. Section 3.5 poses the question of how a knowledge of the behavior at an interface of each component z^n can be synthesized to predict the interaction of a ***general wave packet*** with a boundary. The answer requires solution of an integral equation, and appears to be related to the ***Wiener-Hopf*** technique (but not in the same way as the results of Strang and Osher mentioned in §0.2). This approach is new and, we believe, quite promising, but we do not develop it. Section 3.6 goes on to extend our reflection and transmission results to diagonalizable systems of difference equations. First, interface problems are reduced to boundary problems by a device known as the *folding trick*. This leads to a general ***reflection coefficient matrix*** $[D^{[r]}]^{-1}D^{[\ell]}$ describing reflection and transmission at an arbitrary boundary or interface.

Many of the ideas of Chapter 3 have appeared before, but it is likely that this is the first general description of how to analyze numerical wave behavior at boundaries and interfaces. What makes the general treatment possible is the elimination of any distinction between physical and parasitic waves, and indeed of any reference to the system of equations being modeled, in favor of the notions of leftgoing and rightgoing signals determined by the numerical group velocity.

Chapter 4. In §4 the dissertation turns to ***stability*** for ***initial boundary value problems*** (or interface problems), which we view as a direct outgrowth of reflection and transmission studies. Most of the ideas in this chapter are entirely new. They are however heavily influenced by, and closely tied to, the results of Gustafsson, Kreiss, and Sundström [Gu72]. Section 4.1 begins by explaining the instability of a simple example of an initial boundary value problem model in two ways. First, the *spontaneous rightgoing solution* view considers that the model is unstable because it admits as a solution a set of waves all of which are rightgoing (pointing from the boundary into the field). Second, the *infinite reflection coefficient* view explains instability as the existence for some frequency as a right/left reflection coefficient that is infinite. Sections 4.2-4.3 proceed to analyze mainly the first point of view, which is equivalent to the GKS theory. In §4.2 we first present the *Godunov-Ryabenkii stability criterion* as a statement on strictly rightgoing solutions with $|z| > 1$ (Thm. 4.2.1), and as a *determinant condition* involving the reflection matrices $D^{[r]}$ and $D^{[\ell]}$ (Thm. 4.2.2). Then it is shown that the existence of an arbitrary spontaneous strictly rightgoing solution implies *ℓ_2-instability*, with a growth rate in ℓ_2 proportional to $\sqrt{n}$ (Thm. 4.2.3). We conjecture further that this rate becomes n if an infinite reflection coefficient is present. Thm. 4.2.4 shows that such an unstable solution always causes growth at rate n with respect to boundary data. (Proofs are deferred to Appendix B.) Section 4.3 moves to the stricter *GKS stability definition*, showing by a wave propagation argument why even a non-strictly rightgoing steady-state solution is GKS-unstable (Thms. 4.3.1, 4.3.2). In Section 4.4 the results obtained in §4.1-§4.3 are specialized to the case of dissipative difference models. Section 4.5 applies the main stability results to describe some general classes of unstable difference models in one space dimension, which are extensions of known examples (Thms. 4.5.1-4.5.4). Section 4.6 considers *stability for multidimensional initial boundary value problems*, sketching the relation between instability in this context and solutions with *rightgoing vector group velocities C*, as described in §1.6. An example is described in Thm. 4.6.1.

Chapter 5. Although certain classes of difference models are unambiguously stable or unstable, there are various *borderline cases* for which the situation is less clear. This has always been a source of difficulty in stability theories for initial boundary value problems, and in particular it is responsible for the complexity of the GKS stability definition. Chapter 5 is devoted to a discussion based on numerical experiments of four important classes of borderline cases that are GKS-unstable but stable in some other respects. First, Section 5.2 discusses models that have *finite reflection coefficients*. These are found to be unstable with respect to boundary data,

but in practice nearly stable with respect to initial data, and stable with respect to the introduction of a second boundary. Section 5.3 examines GKS-unstable solutions consisting of rightgoing but *not strictly rightgoing signals*, especially waves with group velocity 0. For this case too, we conclude that instability appears in practice mainly in response to boundary data, and it is weak. In §5.4 we exhibit a class of GKS-unstable problems with both non-strictly rightgoing instabilities and zero reflection coefficients, the *transparent interface anomaly*, and these are ℓ_2-stable. Finally, §5.5 summarizes our views of stability for models of initial boundary value problems in general, and of the GKS theory in particular.

Chapter 6. The last chapter examines stability for problems with *several boundaries or interfaces*, such as might occur in modeling the domain $x \in [0, 1]$, or in mesh refinement, or in composite difference or boundary formulas. This is a natural place to apply wave propagation ideas, because a purely algebraic approach becomes exceedingly complex. We start in §6.2 with one interface, examining known results of Ciment and Tadmor to the effect that dissipativity implies stability. These we extend to more general results in which the notion of *t-dissipativity* introduced in §2 plays a natural part (Thms. 6.2.1,6.2.2). Section 6.3, however, is devoted to proving by a counterexample that no such theorem holds if two or more interfaces are present, contradicting a claim of Oliger [Ol79]. Thus dissipativity is not a strong enough condition to yield stability in general. For an alternative approach, we move on in §6.4 to consider *reflection coefficients at the boundaries*. Thm. 6.4.1 shows that if all reflection coefficients are at most 1 in modulus, then stability for two-boundary problems is guaranteed. We apply this result to duplicate and extend certain results of Beam, Warming, and Yee related to their concept of *P-stability* for two-boundary problems (Thms. 6.4.2,6.4.3). The same reflection coefficient arguments can be applied quite generally, and in §6.5 we consider what growth rates are possible in several important two-boundary or two-interface contexts. The variety of possible growth rates turns out to be considerable, and they are summarized in Table 6.2. These arguments justify, for example, our claim in §5 that GKS-unstable growth will not be converted to exponential growth when a second boundary is introduced unless an infinite reflection coefficient is present. Finally, Section 6.6 discusses very briefly the prospects for problems with *three or more interfaces*.

1. WAVE PROPAGATION IN FINITE DIFFERENCE MODELS

1.1 Dispersion relations and modified equations

Throughout this dissertation we are concerned with the artificial effects introduced when a partial differential equation is approximated by a finite difference scheme. Since these effects appear no matter how elementary the equation under study may be, we will mainly consider as a model the simple ***one-dimensional wave equation,***

$$u_t = au_x, \qquad a \neq 0. \tag{1.1.1}$$

If initial data are specified for $x \in (-\infty, \infty)$,

$$u(x, 0) = f(x), \tag{1.1.2}$$

then the solution to (1.1.1) for all $t \geq 0$ is the translation

$$u(x, t) = f(x + at). \tag{1.1.3}$$

To analyse the behavior of (1.1.1), one may look for Fourier modes

$$u(x, t) = e^{i(\omega t - \xi x)}, \tag{1.1.4}$$

where ω is the (temporal) **frequency** and ξ is the **wave number***. Obviously (1.1.4) will satisfy (1.1.1) if and only if

$$\omega = -a\xi, \tag{1.1.5}$$

a condition known as the **dispersion relation** for (1.1.1). Although standard Fourier *analysis assumes* $\omega, \xi \in \mathbb{R}$, (1.1.5) *holds for arbitrary* $\omega, \xi \in \mathbb{C}$.

*We will be concerned with linear equations only, so it is enough to study complex exponentials. Results for computations in real arithmetic then follow by taking real parts, or equivalently, by adding a complex wave to its conjugate. The use of $e^{-i\xi x}$ rather than $e^{i\xi x}$ in (1.1.4) is designed to make the formulas for phase and group velocity come out without minus signs; see §1.2.

Let (1.1.1) now be modeled by a finite difference formula. For this we set up a regular grid in x and t with spatial step size h, temporal step size k, and **mesh ratio** $\lambda = k/h$, and seek to approximate u by a grid function v:

$$v_j^n \approx u(jh, nk), \qquad j, n \in \mathbb{Z}.$$

One difference formula for (1.1.1) that we will consider repeatedly is **leap frog (LF)**, given by

$$LF: \qquad v_j^{n+1} - v_j^{n-1} = \lambda a(v_{j+1}^n - v_{j-1}^n) \tag{1.1.6}$$

Substituting (1.1.4) into (1.1.6) gives

$$e^{i\omega k} - e^{-i\omega k} = \lambda a(e^{-i\xi h} - e^{i\xi h}),$$

that is,

$$\sin \omega k = -\lambda a \sin \xi h. \tag{1.1.7}$$

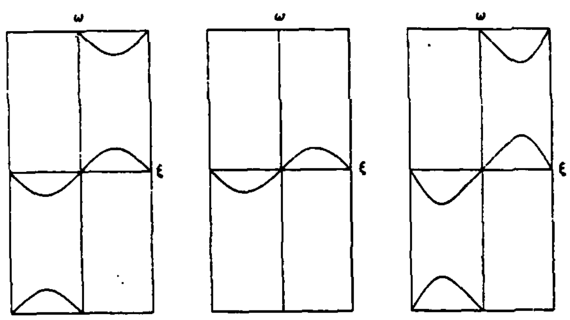
This is the ***dispersion relation for LF***. For small ωk and ξh, which is to say for waves that are well resolved on the grid, (1.1.7) approximates (1.1.5) closely, but as ωk and ξh increase, the approximation becomes poor. Moreover unlike (1.1.5), (1.1.7) is periodic with period 2π in both ξh and ωk. The explanation of this is that because of the discreteness of the grid, any pair $(\xi h, \omega k)$ is indistinguishable on the grid from all of its "aliases" $(\xi h + 2\mu\pi, \omega k + 2\nu\pi)$. Therefore it is enough to consider the fundamental region $(\xi h, \omega k) \in (-\pi, \pi]^2$. Figure 1.1a shows a plot of (1.1.7) in this region for $a = -1$ and $\lambda = .5$. It is apparent that even here, each of ξ or ω corresponds in general to two values of the other variable.*

Solving for ω in (1.1.7), one obtains

$$\omega = \frac{-1}{k} \sin^{-1}(\lambda a \sin \xi h). \tag{1.1.8}$$

By taking the standard branch of the inverse sine here, we confine our attention to the component of the dispersion relation that passes though the origin in Fig. 1.1a.

*The high-frequency lobes of the dispersion curves visible in Fig. 1.1a (and 1.1c) are suggestive of *optical modes* of vibration in crystals, so called because their frequencies are such that they are normally excited by light rather than sound [Bo54]. The physics is quite different, however, for optical modes represent alternative modes of spatial oscillation caused by the presence of multiple species of atoms, whereas the high-frequency components in Fig. 1.1 result from the time discretization.



(a) LF (b) CN (c) LF4

FIG. 1.1. Numerical dispersion relations for difference models LF, CN, and LF4 of $u_t = -u_x$, plotted for mesh ratio $\lambda = .5$. Each plot shows the region $[-\pi/h, \pi/h]^2$ of (ξ, ω)-space. The slope at a point (ξ, ω) is the corresponding group velocity. Additional dispersion plots are given in Appendix A.

Expanding for $\xi h \approx 0$, we get the series

$$\omega = -a\xi\left[1 - \frac{1-(\lambda a)^2}{6}(\xi h)^2 + \frac{1-10(\lambda a)^2+9(\lambda a)^4}{120}(\xi h)^4 + \cdots\right]. \tag{1.1.9}$$

The first term here agrees with (1.1.5), and this must be true for any consistent difference model. From the next term in the series, it is evident that the errors committed by LF will increase with the square of ξh. Formally, (1.1.9) is equivalent to a differential equation of infinite order,

$$u_t = a\left[u_x + \frac{1-(\lambda a)^2}{6}h^2 u_{xxx} + \frac{1-10(\lambda a)^2+9(\lambda a)^4}{120}h^4 u_{xxxxx} - \cdots\right]. \tag{1.1.10}$$

Since (1.1.10) contains derivatives of higher order than 1 but no even-order derivatives, LF is said to be *dispersive* but not *dissipative*. The significance of dispersion is that different wave numbers will travel at different speeds, so that an initial pulse will change shape as time passes. We will examine this in the next few sections. Dissipation will be defined more precisely in §2.2.

As a familiar dissipative scheme, we may consider Lax-Wendroff (LW):

$$LW: \qquad v_j^{n+1} - v_j^n = \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{(\lambda a)^2}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n). \tag{1.1.11}$$

Corresponding to (1.1.7) and (1.1.10), we find for LW the dispersion relation

$$-i(e^{i\omega k} - 1) = -\lambda a \sin \xi h + 2i(\lambda a)^2 \sin^2 \frac{\xi h}{2} \tag{1.1.12}$$

and the formal differential equation of infinite order

$$u_t = a\left[u_x + \frac{1-(\lambda a)^2}{6}h^2 u_{xxx} - \frac{(\lambda a)-(\lambda a)^3}{8}h^3 u_{xxxx} + \frac{1+5(\lambda a)^2-6(\lambda a)^4}{120}h^4 u_{xxxxx} + \frac{(\lambda a)-(\lambda a)^3}{48}h^5 u_{xxxxxx} + \cdots\right]. \tag{1.1.13}$$

It is the non-centered shape of the stencil for LW that gives rise to the complex dispersion relation (1.1.12) and to the even-order derivatives in (1.1.13).

If a difference model is applied to a set of initial data that is smooth in the sense that most of the energy in its Fourier transform has ξh, $\omega k \ll 1$, then one may expect that the model will behave approximately like a differential equation obtained by taking the first few terms of an expansion like (1.1.10) or (1.1.13). This is the idea behind *modified equations* (also known as *model equations*) of difference formulas [Ch83,Wa74]:

Defn. Let a consistent difference model Q of (1.1.1) be formally expanded as a differential equation of infinite order as in (1.1.13). The **modified equation** of Q is the differential equation

$$u_t = a u_x + A h^{\alpha-1} \frac{\partial^\alpha u}{\partial x^\alpha} + B h^{\beta-1} \frac{\partial^\beta u}{\partial x^\beta} \qquad A, B \neq 0, \tag{1.1.14}$$

with α odd and β even, obtained by dropping all but the first dissipative and first dispersive terms from this equation. If there are no dissipative terms we drop the second term and set $\beta = \infty$. //

For example, the modified equation for LW is

$$u_t = a\left[u_x + \frac{1-(\lambda a)^2}{6} h^2 u_{xxx} - \frac{(\lambda a)-(\lambda a)^3}{8} h^3 u_{xxxx}\right]. \tag{1.1.15}$$

We define further

Defn. The integers α and β are the **order of dispersion** and **order of dissipation** of Q. The **order of accuracy** is $\min\{\alpha, \beta\} - 1$. (Consistency implies that the order of accuracy is at least 1.) //

Thus LW, with $\alpha = 3$ and $\beta = 4$, is accurate of order 2, dispersive of order 3, and dissipative of order 4. If $\beta < \alpha$, then dissipation dominates dispersion at low wave numbers, while if $\alpha < \beta$ the reverse holds. We will see in §1.4 that a difference scheme for (1.1.1) is stable in L_p norms, $p \neq 2$, only in the former case.

In this dissertation we will mostly be concerned with nondissipative schemes like LF, because their wave propagation properties are simpler and they are more prone to instabilities. Two other nondissipative models of (1.1.1) that we will often consider are the implicit scheme **Crank-Nicolson (CN)**,

$$CN: \qquad v_j^{n+1} - v_j^n = \frac{\lambda a}{2}\left[\frac{1}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{1}{2}(v_{j+1}^{n+1} - v_{j-1}^{n+1})\right], \tag{1.1.16}$$

and **fourth-order leap frog (LF4)** (fourth order in space, second order in time),

$$LF4: \qquad v^{n+1} - v^{n-1} = \lambda a\left[\frac{4}{3}(v_{j+1}^n - v_{j-1}^n) - \frac{1}{6}(v_{j+2}^n - v_{j-2}^n)\right]. \tag{1.1.17}$$

For CN the dispersion relation is

$$2 \tan \frac{\omega k}{2} = -\lambda a \sin \xi h, \tag{1.1.18}$$

and for LF4 it is

$$\sin \omega k = -\frac{4\lambda a}{3} \sin \xi h + \frac{\lambda a}{6} \sin 2\xi h. \tag{1.1.19}$$

These relations are plotted, again for $a = -1$ and $\lambda = .5$, in Fig. 1.1b–c. One can see that LF4 approximates (1.1.5) better at the origin than LF or CN.

Here are two further examples of dissipative formulas. An implicit formula with $\alpha = 3$, $\beta = 2$ is **backwards Euler (BE)**:

$$BE: \qquad v_j^{n+1} - v_j^n = \frac{\lambda a}{2}(v_{j+1}^{n+1} - v_{j-1}^{n+1}) \tag{1.1.20}$$

An explicit formula with $\alpha = 3$, $\beta = 4$ is **leap frog with dissipation (LFd)** [Kr73, §9],

$$LFd: \qquad v_j^{n+1} - v_j^{n-1} = \lambda a(v_{j+1}^n - v_{j-1}^n) - \frac{\epsilon}{16}(v_{j+2}^{n-1} - 4v_{j+1}^{n-1} + 6v_j^{n-1} - 4v_{j-1}^{n-1} + v_{j-2}^{n-1}), \tag{1.1.21}$$

where $\epsilon \in \mathbb{R}$ lies in the range $0 < \epsilon < 1$.

The properties of the difference schemes we have mentioned are summarised in Appendix A. The Appendix also gives information on several other formulas: Upwind, Box, Method of Lines, Lax-Friedrichs, and Leap Frog for the second-order equation $u_{tt} = a^2 u_{xx}$.

1.2 Phase speed and group speed

Consider now a Fourier mode (1.1.4) in which ω and ξ are both real. It is obvious that in this wave, each point of fixed phase travels at a constant rate

$$c = \frac{\omega}{\xi}, \tag{1.2.1}$$

which is called the **phase speed**. In the case of LF, (1.1.8) and (1.1.9) show that the phase speed is given as a function of ξ by

$$c = \frac{-1}{\xi k} \sin^{-1}(\lambda a \sin \xi h) \approx -a\left[1 - \frac{1-(\lambda a)^2}{6}(\xi h)^2\right]. \tag{1.2.2}$$

Thus LF introduces phase speed errors that increase quadratically as the grid becomes more coarse. Numerical analysts often evaluate difference formulas by examining their phase or phase speed errors (see e.g. §4 of [Ch79b]).

In most applications, however, phase speed is of only secondary importance in determining how an equation behaves. According to a theory initiated by William Hamilton (1839) and Lord Rayleigh (1877), and developed further by Sommerfeld

(1912) and Brillouin, the flow of *energy* in a dispersive medium obeys a **group speed**, defined by

$$C = \frac{d\omega}{d\xi}. \tag{1.2.3}$$

For example, suppose a wave train is formed as a sinusoid with wave number ξ multiplied by a slowly varying envelope $A(x)$. Then as t increases the envelope will move, approximately unchanging in shape, at speed $C(\xi)$, not $c(\xi)$. As a general principle, phase speed controls the interference of waves, but group speed controls their propagation in space.

Eq. (1.2.3) *seems surprising to many people at first, even impossible.* For example one might argue, how can the energy associated with a wave number ξ feel the influence of nearby wave numbers, as (1.2.3) implies that it must? The answer is that polychromatic waves cannot be understood purely in terms of the individual sine waves that make them up—which after all, are each unbounded in extent. It is obvious that the position and structure of any polychromatic pulse are determined by constructive and destructive interference *between* sine waves; so that the "energy associated with wave number ξ", in the absence of other wave numbers, is not localised at all. Therefore it should not be surprising that its propagation with t also depends on the interaction *of wave numbers*. Nevertheless, eq. (1.2.3) takes some getting used to, and readers unfamiliar with group velocity are encouraged to take a look at [Br60], [Wh74], or [Li78].

As a simplest example to motivate (1.2.3), suppose a signal $e^{i(\omega_1 t - \xi_1 x)} + e^{i(\omega_2 t - \xi_2 x)}$ is formed by the superposition of two waves, with $\xi_1 \approx \xi_2$ and $\omega_1 \approx \omega_2$. Then beating will occur. The composite wave is in fact equivalent to a single wave of wave number $(\xi_2 + \xi_1)/2$ modulated by a sinusoidal envelope of wave number $(\xi_2 - \xi_1)/2$, and simple algebra shows that as t increases, the envelope moves at the speed

$$\frac{\omega_2 - \omega_1}{\xi_2 - \xi_1}.$$

This approaches (1.2.3) in the limit $\xi_2 \to \xi_1, \omega_2 \to \omega_1$.

A more general derivation of group velocity is based on the ***method of stationary phase***, due to Lord Kelvin. (For further derivations, see [Wh74] and [Li78], and also §2.3.) Let an initial distribution $u(x,0) = f(x)$ have the Fourier transform $\hat{f}(\xi)$. Let this signal propagate with t according to a dispersion function $\omega = \omega(\xi)$.* Then at

*For a treatment of a multivalued dispersion relation, as is needed for multilevel difference schemes, see Appendix B.

time $t \geq 0$, the solution (ignoring normalization factors) is

$$\begin{aligned} u(x,t) &= \int e^{i(\omega(\xi)t - \xi x)} \hat{f}(\xi)\, d\xi \\ &= \int e^{it(\omega(\xi) - \xi x/t)} \hat{f}(\xi)\, d\xi. \end{aligned} \tag{1.2.4}$$

Suppose x/t is held fixed as $t \to \infty$. This corresponds to moving our eyes rightward at a fixed speed x/t =const. After a long time, what will we see? The answer comes from observing that as t increases, the exponential in (1.2.4) oscillates more and more rapidly with ξ, hence tends to cancel to 0 as $t \to \infty$. Assuming that $\hat{f}$ is smooth enough, which will be the case if f is localized, such cancellation will evidently take place everywhere except for any ξ of *stationary phase*, at which

$$\frac{d}{d\xi}(\omega - \xi x/t) = 0,$$

i.e.

$$\frac{d\omega}{d\xi} = \frac{x}{t}.$$

As $t \to \infty$, therefore, our eyes will see only any wave numbers that satisfy this equation. In other words, energy associated with wave number ξ moves asymptotically at the group speed (1.2.3).

The stationary phase argument is made quantitative in [Br60], [Li78], and [Wh74]. In App. B (Lemma B.1), we will give a complete argument of a related kind in order *to prove* the stability theorems of Chapter 4.

Since the stationary phase idea is applicable in various contexts, we have left out details such as limits of integration, but let us now be more precise for the problem of central interest. If f is a discrete function defined only for $x = jh$, $j \in \mathbb{Z}$, then $\hat{f}$ is defined by a infinite sum and has domain $[-\pi/h, \pi/h]$, so the limits of integration in (1.2.4) become $\pm\pi/h$. For $f \in \ell_2(h)$, one has $\hat{f} \in L_2[-\pi/h, \pi/h]$, and the more localized f is, the smoother $\hat{f}$ will be; when f has compact support, $\hat{f}$ will be a trigonometric polynomial. Whether or not f has compact support, its domain can be extended naturally from $h\mathbb{Z}$ to all of $\mathbb{R}$ by simply evaluating (1.2.4) for arbitrary x. The result is a function in $L_2(-\infty, \infty)$, namely the (finite or infinite) trigonometric interpolant through the values $\{f(jh)\}$. By Parseval's formula, the L_2 norm of this extension will equal the ℓ_2 norm of the discrete function f (if both are appropriately normalized), since both are equal to the L_2 norm of $\hat{f}$. Therefore in later sections we can study the sum-of-squares energy of a signal without being too careful as to whether we consider its domain to be continuous or discrete.

Now let us examine the group speed for waves under LF. By differentiating (1.1.7) implicitly on both sides, one obtains

$$k \cos \omega k \, d\omega = -a\lambda h \cos \xi h \, d\xi,$$

hence

$$C = -a \frac{\cos \xi h}{\cos \omega k}. \tag{1.2.5}$$

This formula shows that the effects of discretization in x and t multiply each other; for small ξh and ωk the former will tend to decrease $|C|$ and the latter to increase it (cf. §2.1). Since stability requires $\lambda|a| < 1$, the first effect will dominate. By (1.1.8), we can eliminate ωk to get

$$C = \frac{-a \cos \xi h}{\sqrt{1 - (\lambda a)^2 \sin^2 \xi h}} \approx -a\Big[1 - \frac{1 - (\lambda a)^2}{2}(\xi h)^2\Big]. \tag{1.2.6}$$

A comparison of (1.2.2) and (1.2.6) shows that for small ξh and ωk, both c and C will be less than the ideal speed $-a$ in magnitude, but that C will lag by roughly three times as much.

Similarly, differentiating (1.1.18) leads to the group speed

$$C = -a \cos \xi h \, \cos^2 \frac{\omega k}{2} \approx -a\Big[1 - \frac{2 + \lambda a^2}{4}(\xi h)^2\Big] \tag{1.2.7}$$

for CN, and (1.1.19) gives

$$C = -a \frac{\frac{4}{3} \cos \xi h - \frac{1}{3} \cos 2\xi h}{\cos \omega k} \approx -a\Big[1 + \frac{\lambda a^2}{2}(\xi h)^2\Big] \tag{1.2.8}$$

for LF4. Since $C = d\omega/d\xi$, these functions represent the slopes of the dispersion relation plots in Fig. 1.1.

From these formulas one can calculate that with LF4 and CN as with LF, C lags the ideal value for $\xi h \approx 0$ by 3 times as much as c. This fact generalizes as follows:

Theorem 1.2.1. *Let Q be a nondissipative model of $u_t = au_x$ with the modified equation*

$$u_t = au_x + Ah^{\alpha-1} \frac{\partial^\alpha u}{\partial x^\alpha} \tag{1.2.9}$$

for some odd integer $\alpha \geq 3$. Then as ξh, $\omega k \to 0$, the phase and group speeds satisfy

$$c = -a - (-1)^{\frac{\alpha-1}{2}} A(\xi h)^{\alpha-1} + O((\xi h)^\alpha),$$

$$C = -a - \alpha(-1)^{\frac{\alpha-1}{2}} A(\xi h)^{\alpha-1} + O((\xi h)^\alpha).$$

Thus C differs from the ideal speed $-a$ by α times as much as c.

Proof. Eq. (1.2.9) implies that for small ξh, ωk the dispersion relation is

$$i\omega = -ia\xi + Ah^{\alpha-1}(-i\xi)^\alpha,$$

i.e.

$$\omega = -a\xi - (-1)^{\frac{\alpha-1}{2}} Ah^{\alpha-1}\xi^\alpha.$$

The result now follows from (1.2.1) and (1.2.3). ∎

This theorem implies that *evaluation of difference formulas by the phase errors they introduce may lead to unrealistically optimistic conclusions.*

DEMONSTRATION 1.1. As the simplest demonstration of group speed, Fig. 1.2 shows the propagation of a nearly monochromatic wave packet under LF with $a = -1$, $\lambda = .4$. Fig. 1.2a plots the initial signal on a grid with $h = 1/160$,

$$u(x, 0) = e^{-(8x)^2} \sin \xi x,$$

with ξ chosen so that there are 8 grid points per wavelength: $\xi h = 2\pi/8 \approx .79$, $\xi \approx 251.3$. The exact solution should move right unchanged at speed 1, but (1.2.2) and (1.2.6) predict phase and group speeds

$$c \approx .91, \qquad C \approx .74.$$

In this experiment the exact solution was used to provide values at $t = k$, and then LF was applied up to $t = 1$. The result is shown in Fig. 1.2b. Apparently the wave packet has propagated at just the group speed C, not at the phase speed, and it has changed little in shape. If one looked at the wave carefully as a function of t, one would see phase crests continually appearing at the trailing edge of the packet, advancing through it at speed c, and disappearing at the front. The same behavior appears in the ripples made when a stone is dropped into a pond, for gravity waves

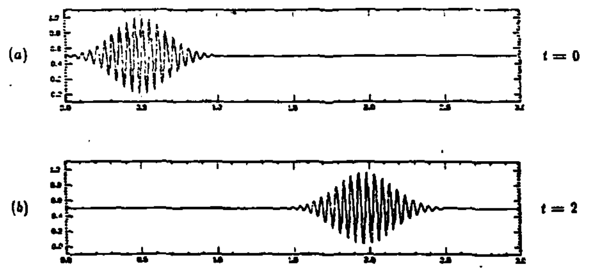


FIG. 1.2. Propagation of a wave packet with 8 points per wavelength ($\xi h \approx .79$). The model is LF for $u_t = -u_x$ with $h = 1/160$, $\lambda = .4$. The packet moves not at the ideal speed 1, but at the group speed $C \approx .74$.

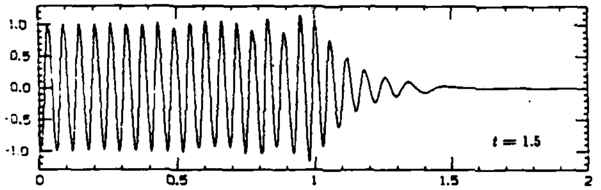


FIG. 1.3. Propagation of a wave front generated by a forced oscillation with $\omega k = 1$ at the left boundary. The model is CN for $u_t = -u_x$ with $h = 1/500$, $\lambda = 5$. The wave front travels at the group speed $C \approx .75$.

on deep water also satisfy a dispersive equation characterized by $C < c$.*

This example demonstrates a principle that makes analysis of group velocity errors in difference schemes possible: *there is more to the inaccuracy of a difference scheme than truncation error.* The wave in Fig. 1.2b differs completely from the correct solution pointwise, and so an estimate of accumulated truncation error would lead to the conclusion that the computation had been useless. But in fact, it has been qualitatively correct. *Errors caused by differencing are not random perturbations, but a systematic interaction of dispersions and possibly dissipations of various orders.*

DEMONSTRATION 1.2. As a second example, Fig. 1.3 shows the propagation of a wave front. In this experiment a sinusoidal forcing oscillation at the left boundary radiates a wave into the interior of the interval $[0,2]$. Here $h = 1/500$, $\lambda = 5$, and the scheme is CN with $a = -1$. The oscillation

$$u(0,t) = \sin 100t$$

has been turned on at $t = 0$. At $t = 1.5$, only a low-frequency forerunner has reached $x = 1.5$; the main oscillation of amplitude 1 has reached only $x = 1.0$ or 1.1, suggesting that the wave front propagates at a speed roughly 0.7. Now to analyze a problem like this we need to know how C depends on ω, not ξ. From (1.1.18) and (1.2.7), we obtain

$$C(\omega) = -a\cos^2\frac{\omega k}{2}\sqrt{1-\left(\frac{2}{\lambda a}\right)^2\tan^2\frac{\omega k}{2}}. \tag{1.2.10}$$

For the given problem $\omega k = 1$, and (1.2.10) predicts $C \approx .75$. This explains Fig. 1.3. Throughout this dissertation, we will use both spatial and temporal Fourier transforms as convenient; most often it will be the latter, since boundaries or interfaces will be present.

For dissipative modes, the concept of group velocity breaks down. When dispersion dominates dissipation, the predictions obtained by ignoring dissipation may

*In fact for such waves one has $C = \frac{1}{2}c$. For short ripples on deep water (surface tension dominated), on the other hand, one has $C = \frac{3}{2}c$. Other physical problems with $C > c$ are wave propagation in elastic beams ($C = 2c$) and movement of a particle when viewed as a quantum mechanical wave packet ($C = 2c$ also). (The classical particle speed corresponds to C, not c.) Closer physical analogs to a finite difference model of (1.1.1) are presented by problems in which $C \approx c$ for long wavelengths but $C \neq c$ for short ones. These include sound or electromagnetic wave propagation in random media (air, glass, rock) or regular media (crystals, electric networks). In these cases c and C begin to differ when the wavelengths present become comparable to some physical scale involved, such as a distance between molecules.

not be far off, and we will make use of this in §1.4. One justification of this claim can be found in Thm. 2.3.1 together with Lemma B.1; in fact, Thm. 1.2.1 could be extended to even-order dissipative difference approximations. However, a general analysis requires a steepest descent argument that is more subtle than the stationary phase derivation [Br60]. It turns out that for dissipative waves one can distinguish *group*, *signal*, and *energy* velocities, all of which coincide in the nondissipative case. This theory was worked out by Brillouin and Sommerfeld in the early 1900's and is described at length in [Br60]. The application of steepest descent analysis to dissipative finite difference models of (1.1.1) is carried out by Serdjukova in [Se63,Se66], and by Hedstrom and Chin in [He65,He66,He68,He75,Ch75,Ch78]. The same approach has been extended to models of a transport equation by Gropp [Gr81].

1.3 Dispersion

In a signal consisting of a superposition of various wave parameter pairs (ξ, ω), the energy associated with each pair will propagate at the group speed appropriate to that pair. In general these group speeds will be different, causing the signal to change shape as it propagates. This separation of wave numbers is called **dispersion**.

DEMONSTRATION 1.3. The simplest configuration that may lead to dispersion is a superposition of two wave numbers, a *dichromatic wave packet*. Figure 1.4a shows such a signal, given by

$$u(x,0) = \frac{1}{2}e^{-80(x-1/2)^2}(1 + \sin 100x).$$

This signal contains equal amounts of energy at wave numbers $\xi \approx 0$ and $\xi \approx 100$. In the experiment the LF formula was applied with $a = -1$, $h = 1/100$, $\lambda = .5$, and the exact solution was used to provide data at $t = k$. For these values (1.2.8) predicts that the low wave number energy should move at speed $C = 1$, and the high wave number energy at $C \approx .60$. Figures 1.4b,c show the computed result at $t = 2, 4$. The initial packet has split into two pieces, and they have evidently traveled at the predicted speeds. (Compare Fig. 1 of [Vi75].)

More generally, any wave packet that is localized in space must contain a range of wave numbers. Quantitatively, the product of the width of a wave packet u and the width of its Fourier transform $\hat{u}$ is bounded from below by a constant of order unity (the uncertainty principle). In particular, the initial signal in Fig. 1.4 is not

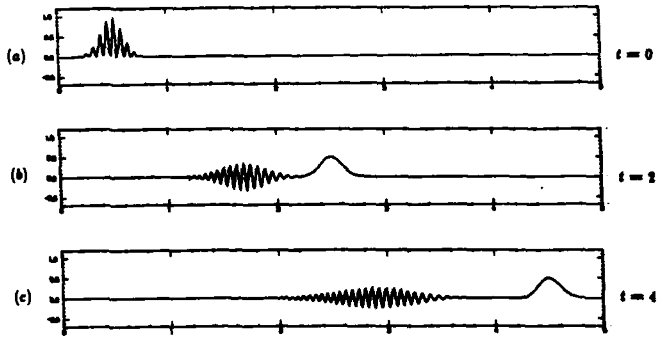


FIG. 1.4. Separation of a dichromatic wave packet with $\xi h \approx 0$ and $\xi h \approx 1$ into two components. The model is LF for $u_t = -u_x$ with $h = 1/100$, $\lambda = .5$.

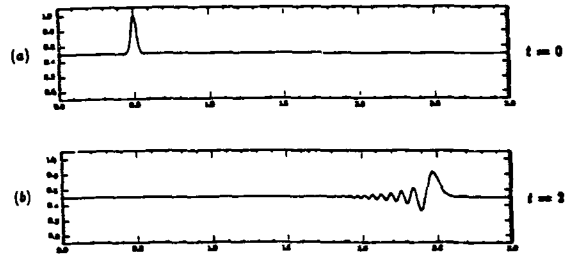


FIG. 1.5. Dispersion of a polychromatic pulse. The model is LF for $u_t = -u_x$ with $h = 1/160$, $\lambda = .4$. Higher wave numbers have lower group speeds and lag behind the main signal.

exactly dichromatic, but has a Fourier transform consisting of two narrow spikes. Similarly the signal in Fig. 1.2 has a spectrum consisting of one narrow spike. In such cases we must expect that each not-quite-monochromatic wave component present will itself disperse with time, since it contains energy with various group velocities. Such dispersion will take the form of a broadening of the wave packet at a steady rate depending on the range of group speeds present. We can formulate this in an approximate way as follows:

Let an initial wave packet $u(x,0)$ have Fourier transform $\hat{u}(\xi,0)$ with support $[\xi_0-\Delta\xi/2, \xi_0+\Delta\xi/2]$ for some small value $\Delta\xi$. Let $W(t)$ be an approximate measure of the width of the packet at time t. Then for large t, W will grow roughly according to

$$W(t) - W(0) \approx t\,\Delta\xi\,\frac{dC}{d\xi}(\xi_0). \qquad (1.3.1)$$

The significance of (1.3.1) is twofold. First, broadening of a pulse will be approximately linear. Second, the rate of broadening depends on both the width of the Fourier transform and the derivative $dC/d\xi$.

In Fig. 1.2, there were many grid points in the wave packet, so $\Delta\xi$ was small and the packet broadened only about 10% or so in the time shown. This example illustrates a point of practical importance: the absence of conspicuous dispersion is no guarantee that a computation has been accurate. In Fig. 1.4 there are not as many grid points within the wave packet, so $\Delta\xi$ is large. The component with $\xi \approx 0$ still does not broaden much, because $dC/d\xi = 0$ at $\xi = 0$. But it is evident that the component with $\xi h = 1$ has broadened considerably. In fact, for this component (1.3.1) has the approximate form

$$W(t) - W(0) = t(20)(\frac{1}{50}).$$

This leads to estimates like

$$W(0) = 0.2, \qquad W(2) = .8,$$

which are not far off, considering that we have been careless with constants.

DEMONSTRATION 1.4. Figure 1.5 shows the dispersion of an initial pulse that is so narrow as to be thoroughly polychromatic (cf. [Vi82]). This experiment takes place in the same laboratory as Demo. 1.1: $a = -1$, $h = 1/160$, $\lambda = 0.4$, scheme = LF.

But the initial distribution is now

$$u(x,0) = e^{-3200(x-\frac{1}{2})^2},$$

which is much narrower than before and has central wave number $\xi = 0$. Since the pulse is narrow, its transform is broad, and Fig. 1.5b shows that it disperses quickly into a train of oscillations.

Such oscillatory effects of finite difference schemes are common and well known. What is not generally recognised is that all of the behavior of Fig. 1.5, except for the phases of individual wave crests, can be predicted quantitatively by considering group speed. At the front of the wave train, the low wave numbers travel at speed nearly 1, as they must. The further back one looks, the higher the wave number one sees; measurements in an enlargement of Fig. 1.5b confirm that the relationship is that of (1.2.8). Furthermore, the amplitude distribution can be predicted from the fact that the initial ℓ_2 energy density at each wave number is conserved (it must first be defined carefully, since LF is a multilevel scheme; see [Ri67]). Accordingly, the amplitude of a part of the wave train with wave number ξ decreases with time according to the square root of the rate of dispersion $dC/d\xi$. These ideas are made precise and applied extensively in the field of *geometrical optics* [Wh74].

For analyses of the dispersion introduced by finite difference models in the neighborhood of a discontinuity in u, see [Ap68,Ch75,He75,Ch78].

1.4 Instability in L_p norms, $p \neq 2$

In this section we digress briefly to consider our first application of wave propagation ideas to stability. We will show that dispersion is the controlling factor for stability of difference models of $u_t = au_x$ in L_p norms, $p \neq 2$.

In the last two decades a considerable body of results has accumulated on stability in L_p norms [Br75]. Some of the contributors to this work have been Brenner, Hedstrom, Serdjukova, Stetter, Thomée, and Wahlbin. This theory is quite technical, and does not draw explicitly on the notions of group velocity or dispersion.* Instead it is founded mainly on the techniques of *Fourier multipliers*. Our contention is that many of these results can be readily understood, and possibly extended, by simpler

*However, G. Hedstrom at least (private communication) has been aware of the interpretation of L_p instability presented here.

arguments. We will only sketch some ideas here without developing them rigorously, as this dissertation is mainly concerned with stability for problems containing boundaries or interfaces. However, the discussion should suffice to provide support for our underlying thesis: that the stability of finite difference models is strongly affected by phenomena of dispersive wave propagation.

Let Q denote a fixed finite difference approximation to $u_t = au_x$ with time and space steps k and $h = k/\lambda$, $\lambda = \text{const}$. We will apply Q at all points $x \in (-\infty, \infty)$ but at discrete time levels nk, and denote the computed solution at time step n by $v^n(x)$. For simplicity we take Q to be a two-level formula, and let S denote the solution operator $v^n \mapsto v^{n+1}$:

$$v^n(x) = [S^n v^0](x). \tag{1.4.1}$$

For $1 \leq p < \infty$ the L_p norm of a function $v : \mathbb{R} \to \mathbb{C}$ is defined by

$$\|v\|_p^p = \int_{-\infty}^{\infty} |v(x)|^p \, dx \qquad (1 \leq p < \infty), \tag{1.4.2}$$

and for $p = \infty$,

$$\|v\|_\infty = \sup_{x \in R} |v(x)|.$$

The space L_p consists of those functions v for which this number is finite. If $S : L_p \to L_p$ is a bounded operator, the induced operator p-norm is given by

$$\|S\|_p = \sup_{\|v\|_p = 1} \|Sv\|_p \qquad (1 \leq p \leq \infty).$$

We define stability in L_p as follows:

Defn. *The model Q is L_p-stable if for each $T > 0$ there exists a constant C_T such that*

$$\|S^n\|_p \leq C_T$$

for all n and k satisfying $nk \leq T$.

For models of hyperbolic problems the L_2 norm is most often used, mainly because it is naturally connected to the Fourier transform by Parseval's formula. But other L_p norms also come up sometimes, particularly the L_1 and L_∞ norms when one has in mind an extension to a nonlinear problem [Lu81]. One might expect that most difference formulas that are stable in L_2 would be stable in other L_p norms too. However, a result due to Thomée shows that this is not so (see p. 100 of [Ri67]):

Theorem 1.4.1 *[Th65]. Let Q approximate $u_t = au_x$ to an even order of accuracy. Then Q is unstable in L_p for all $p \in [1, \infty]$, $p \neq 2$.* ∎

It is this and related results that we claim are due to dispersion.

Here is the explanation. Consider as an initial distribution a narrow pulse, as in Fig. 1.5a, whose width is a few grid points. Following (1.3.1), we write this in the form

$$W(0) \approx h, \tag{1.4.3}$$

with the understanding that "$\approx$" denotes an order of magnitude agreement, ignoring constant factors, without being defined precisely. As n increases, the pulse will disperse into a train of oscillations (Fig. 1.5b), whose width will increase roughly linearly with n (cf. (1.3.1)),

$$W(n) \approx W(0) + t \approx nk. \tag{1.4.4}$$

Let $A(n)$ be some measure of the average amplitude of the wave train. Then we expect to have

$$\|v^n\|_p \approx A(n)[W(n)]^{1/p}. \tag{1.4.5}$$

Now if Q is nondissipative, $\|v^n\|_2$ will be approximately conserved as n increases (exactly, if Q is a two-level formula). With (1.4.3)–(1.4.5), this implies

$$\frac{A(n)}{A(0)} \approx \left[\frac{W(n)}{W(0)}\right]^{-\frac{1}{2}} \approx n^{-\frac{1}{2}}. \tag{1.4.6}$$

Therefore by (1.4.3)–(1.4.6) we have

$$\frac{\|v^n\|_p}{\|v^0\|_p} \approx \frac{A(n)}{A(0)} \left[\frac{W(n)}{W(0)}\right]^{\frac{1}{p}} \approx n^{\frac{1}{p} - \frac{1}{2}}. \tag{1.4.7}$$

For $p < 2$, the exponent is positive, and so we have growth in the p norm. It follows that the operator powers S^n must grow at least this fast,

$$\|S^n\|_p \gtrsim n^{\frac{1}{p} - \frac{1}{2}}, \tag{1.4.8}$$

and since $n \to \infty$ as $k \to 0$ for fixed $t = nk$, this contradicts the definition of L_p-stability. *Therefore Q is unstable in L_p for $p < 2$.*

Thus L_p instability for $p < 2$ can be explained by the dispersion of narrow spikes into *oscillatory wave trains*. Correspondingly, instability for $p > 2$ is implied by the fact that an oscillatory wave train may coalesce into a spike. Suppose that the configuration of Fig. 1.5b is taken as initial data $v^0(x)$, and then the LF model of (1.1.1) is applied with $a = 1$ instead of $a = -1$. (Alternately, one might retain $a = -1$ but reflect the wave train of Fig. 1.5b about $x = 0$.) Then as t increases the wave train will move left, and the lower wave numbers to the right will overtake the higher ones to the left, whose group speeds are not quite as large. The result at $t = 2$ will be another spike at $x = 0$—not identical to that of Fig. 1.5a, but close. From $t = 0$ to $t = 2$, each L_p norm with $p > 2$ will have grown. Now $W(0)$ and $W(n)$ are the same as before except reversed, hence $A(0)$ and $A(n)$ also, and (1.4.7) becomes

$$\frac{\|v^n\|_p}{\|v^0\|_p} \approx n^{\frac{1}{p}-\frac{1}{2}}. \tag{1.4.9}$$

This time the exponent is negative for $p > 2$, and (1.4.8) becomes

$$\|S^n\|_p \gtrsim n^{\frac{1}{p}-\frac{1}{2}}. \tag{1.4.10}$$

Eqs. (1.4.8) and (1.4.10) combine to give the general bound

$$\|S^n\|_p \gtrsim n^{|\frac{1}{2}-\frac{1}{p}|}. \tag{1.4.11}$$

(Actually, for the above argument to go through we must be a little more careful. The problem is that the wave train of Fig. 1.5b is not at all uniform in amplitude, so that $A(n)$ cannot be defined in such a way that (1.4.5) holds for all p. The explanation for this comes from (1.3.1) and the discussion in §1.3: our initial spike contains both nonzero wave numbers, which broaden and therefore decay in amplitude because they have $dC/d\xi \not\approx 0$, and near-zero ones, which decay very little because they have $dC/d\xi \approx 0$. One remedy is to replace Fig. 1.5a with a signal that looks more like the derivative of a spike. The Fourier transform of the proper signal, instead of being concentrated in a band of width $\Delta\xi$ at $\xi = 0$, might consist of a band of width $\Delta\xi$ centered at $\xi = \Delta\xi$. Then the broadening rates of the various energy components, hence their amplitude decay rates too, will agree up to constant factors, and (1.4.5) will be valid.)

Now suppose Q is dissipative. Here is the explanation for the even-order hypothesis of Thm. 1.4.1. If Q has even order of accuracy, then its model equation has $\alpha < \beta$ (§1.1), and this means that dispersion is stronger than dissipation at low wave numbers. By considering a spike as before composed of energy with sufficiently low wave numbers, we can again get growth in all L_p norms, $p \neq 2$. On the other hand

if Q has odd order of accuracy, then dissipation dominates dispersion, and we cannot achieve such growth.

Let us substantiate these claims by estimating the growth rate for an even-order formula with $\alpha < \beta < \infty$. In the nondissipative case, we took an initial signal with width $W(0) \approx h$. The trouble is, the transform of such a signal is so broad that the energy will tend to dissipate faster than it disperses. On the other hand if $W(0)$ is taken too large, then although the dissipation is small, we will have a wide packet broadening slowly, and not much growth will take place. Achieving a maximum growth rate will depend on picking $W(0)$ so as to balance these effects. $W(0)$ will also have to depend on what time step n it is at which we wish to observe growth. The reason is that the growth due to dispersion is algebraic, while the decay due to dissipation is exponential; for large enough n, the simple kind of packet we are considering will decay to 0 in all p norms.

The maximal growth solution is this: given n, design an initial packet as before but with

$$W(0) \approx hn^{\frac{1}{\beta}}. \tag{1.4.12}$$

The width of the Fourier transform is then

$$\Delta\xi \approx h^{-1}n^{-\frac{1}{\beta}}. \tag{1.4.13}$$

If the order of dispersion is α, a packet of this width will have group velocities covering a range (Thm. 1.2.1)

$$\Delta C \approx (h\Delta\xi)^{\alpha-1} \approx n^{\frac{1-\alpha}{\beta}},$$

and so W will increase with n according to

$$W(n) \approx W(0) + t\,\Delta C \approx hn^{\frac{\beta+1-\alpha}{\beta}}. \tag{1.4.14}$$

Eqs. (1.4.12) and (1.4.14) give the ratio of widths

$$\frac{W(n)}{W(0)} \approx n^{\frac{\beta-\alpha}{\beta}}. \tag{1.4.15}$$

To get the corresponding ratio of amplitudes, we observe that since Q has order of dissipation β, the L_2 norm of v will decay according to

$$\|v\|_2 \approx (1-(h\Delta\xi)^\beta)^n \approx e^{-n(h\Delta\xi)^\beta},$$

or by (1.4.13),

$$\|v^n\|_2 \approx e^{-1} \approx 1.$$

In other words, our initial packet is just broad enough so that the decay up to step n is not significant. (The width (1.4.12) was chosen to be the smallest possible for which this would hold.) Therefore as in (1.4.6) we have by (1.4.15),

$$\frac{A(n)}{A(0)} \approx \left[\frac{W(n)}{W(0)}\right]^{-\frac{1}{2}} \approx n^{\frac{\alpha-\beta}{2\beta}}. \tag{1.4.16}$$

From this follows the analog of (1.4.7),

$$\frac{\|v^n\|_p}{\|v^0\|_p} \approx \frac{A(n)}{A(0)}\left[\frac{W(n)}{W(0)}\right]^{\frac{1}{p}} \approx n^{\frac{\alpha-\beta}{2\beta}} n^{\frac{1}{p}\left(\frac{\beta-\alpha}{\beta}\right)} = n^{\frac{\beta-\alpha}{2\beta}\left(\frac{2}{p}-1\right)},$$

or following (1.4.8),

$$\|S^n\|_p \gtrsim n^{\frac{\beta-\alpha}{2\beta}\left(\frac{2}{p}-1\right)}. \tag{1.4.17}$$

For $p < 2$ the exponent is again positive. *Therefore Q is unstable in L_p for $p < 2$.*

As before, reversing the process gives the same estimate but with the exponent negated, implying growth in L_p for $p > 2$. All together, we have the bound*

$$\|S^n\|_p \gtrsim n^{\frac{\beta-\alpha}{2\beta}\left|\frac{2}{p}-1\right|}. \tag{1.4.18}$$

This agrees with the nondissipative result (1.4.11) if one sets $\beta = \infty$.

The above arguments constitute a sketch of a proof of Thm. 1.4.1.

In addition, we have obtained a lower bound for the growth rate of the difference operator. What is remarkable is that this bound is as strong as possible. The following result was proved by Brenner, Thomée, and Wahlbin:

Theorem 1.4.2 *[Br75, Thms. 3.1,3.2]. Let Q be a consistent difference approximation to $u_t = au_x$ with even order of accuracy. If Q is dissipative, the powers of the solution operator S satisfy for $1 \le p \le \infty$ a bound*

$$M_1 n^{\frac{\beta-\alpha}{\beta}\left|\frac{1}{2}-\frac{1}{p}\right|} \le \|S^n\|_p \le M_2 n^{\frac{\beta-\alpha}{\beta}\left|\frac{1}{2}-\frac{1}{p}\right|} \tag{1.4.19}$$

for some constants M_1 and M_2. If Q is nondissipative ($\beta = \infty$) this reduces to the formula

$$M_1 n^{\left|\frac{1}{2}-\frac{1}{p}\right|} \le \|S^n\|_p \le M_2 n^{\left|\frac{1}{2}-\frac{1}{p}\right|}. \quad \blacksquare \tag{1.4.20}$$

*An application of the uniform boundedness principle shows that not only does S^n grow at this rate as $n \to \infty$, but so does v^n for some suitably chosen initial data v^0. In fact it is not hard to devise such a function v^0: let it consist of a series of spikes, each broader and weaker than the last, and each designed to achieve maximum growth at a particular time step.

The fact that our estimate was sharp suggests that not only does the dispersion and gathering of spikes imply instability in L_p norms, but there is nothing more to such instability than this.

Much of the early research leading to Thm. 1.4.2 was concerned with growth rates in L_∞ of the Lax-Wendroff operator, $Q = \mathrm{LW}$. For this we have $\alpha = 3, \beta = 4, p = \infty$, and (1.4.19) becomes

$$M_1 n^{1/8} \le \|S^n\|_\infty \le M_2 n^{1/8}. \tag{1.4.21}$$

(Obviously the instability is very weak.) This bound was first established by Serdjukova [Se63], by means of saddle-point estimates, and independently by Hedstrom [He66] in 1966; see also [St65], [Th65], [Se66].

The theory related to instability in L_p has been carried well beyond Thm. 1.4.2. In particular one may ask, how rapidly can $\|v^n\|_p$ grow if v^0 satisfies some smoothness condition? How smooth must v^0 be to make growth impossible? The answers to such questions naturally involve *Besov spaces* [Pe76], and a large number are presented in [Br75]; see also [He68]. Although many of them could probably be given dispersion interpretations, we do not argue that this would necessarily be productive in the more complicated cases.

A more promising application of the dispersion idea may involve the extension to *variable coefficients*. We have not discussed this fact, but the essentials of group velocity extend without change to dispersive systems with variable coefficients, so long as the scale of variation is large compared to the wavelengths of interest [Wh74]. Therefore we propose:

Conjecture. *Theorem 1.4.1 continues to hold if Q is a consistent difference model of*

$$u_t = a(x)u_x,$$

where $a(x)$ is a Lipschitz continuous function satisfying $0 < a_{\min} \le |a(x)|$ for all x.

A straightforward extension of the estimate (1.4.19) is probably also valid. At present no such results appear to have been proved, although some theorems for variable coefficients appear in [Ap68]. To apply Fourier methods, one would most likely need to move from Fourier multiplier techniques to those of *pseudodifferential operators*. Technically this would be intricate, and there is a chance that the resulting theorems would require an unreasonable degree of smoothness in a. We suspect that a proof by arguments based on dispersion would be easier to carry out.

For some very interesting results on stability in L_p norms of nonlinear difference formulas, see the recent dissertation by B. Lucier [Lu81].

1.5 Parasitic waves

The last three sections have concentrated on the errors that result from the deviation from linearity of a numerical dispersion relation near the origin $\omega = \xi = 0$. These might be called the *behavioral errors* introduced by differencing. However, a finite difference grid can also support completely nonphysical or **parasitic waves**, with ξh or ωk far from the origin, and these too will propagate at the group speed (1.2.3). In general parasitic waves may travel not only at the wrong speed, but also in the *wrong direction*. This can be seen from the fact that in Fig 1.1 (also Appendix A), the dispersion curves have negative slope in various regions. In Chapter 4 we will see that energy propagation in the wrong direction is closely related to instability for initial boundary value problems.

It is perhaps surprising that poorly resolved waves should obey a group speed, since the discreteness of the grid might seem to necessitate a more complicated analysis. However, the stationary phase argument sketched in §1.2 only required $\hat{u}(\xi, 0)$ to be smooth function of ξ, and has nothing to do with the discreteness of x.

DEMONSTRATION 1.5. *To illustrate*, Fig. 1.6 shows the propagation of five different wave packets. In this experiment $u_t = -u_x$ with $a = -1$ was modeled on $[-1.5, 1.5]$ by CN with $\lambda = 1$, $h = 1/100$. In each case initial data consisted of a wave packet

$$v^0(x) = e^{-(10x)^2} \cos \xi x,$$

with varying values of ξ. In each case the solution was computed up to $t = 1$, and then the result was plotted. From (1.1.18) and (1.2.7), one readily obtains the prediction

$$C = \frac{\cos \xi h}{1 + \frac{1}{4}\sin^2 \xi h}.$$

for this demonstration. Table 1.1 shows the wave numbers used and the corresponding group speed predictions:

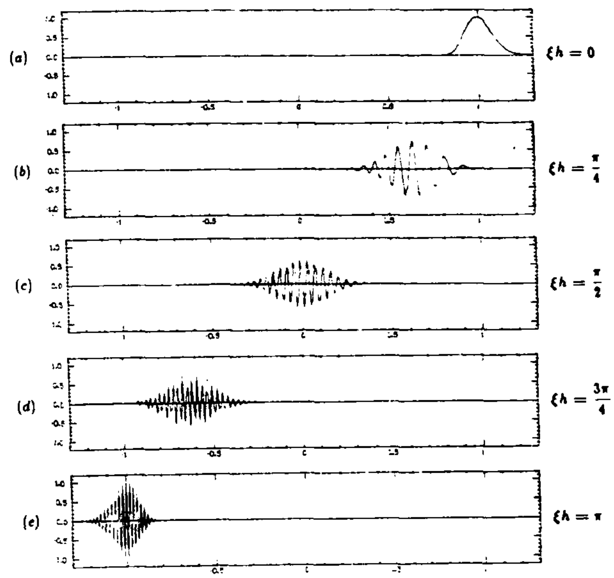


FIG 1.6. Physical and parasitic wave packets with $\xi h = 0, \pi/4, \ldots, \pi$. In each experiment the initial packet was located at $x = 0$ and the figure shows the result at $t = 1$, so that the position of each packet plotted equals the group velocity for the corresponding wave number. The model is CN for $u_t = -u_x$ with $h = 1/100$, $\lambda = 1$.

Figure	ξh	C
1.6a	0	1
1.6b	$\pi/4$	.629
1.6c	$\pi/2$	0
1.6d	$3\pi/4$	$-.629$
1.6e	π	-1

TABLE 1.1

The figure shows clearly that the predictions of this table are valid.

DEMONSTRATION 1.6. Figure 1.7 shows the similarity between physical waves and parasites in another way. In addition to the spatially sawtoothed waves that we have already seen, which arise from near $(\xi,\omega) = (\pi/h, 0)$, Figs. 1.1a–c imply that signals with (ξ,ω) near $(0,\pi/k)$ and $(\pi/h,\pi/k)$ are also possible under LF or LF4. Fig. 1.7 confirms this for the scheme LF with $a = -1$. In the same mesh as before, sinusoidal forcing functions with $\omega k = 0, 0.1, \pi$ have now been turned on at $t = 0$ in the *middle* of the domain:

(1.7a) $v_0^n = 1$,

(1.7b) $v_0^n = \sin(.1n)$,

(1.7c) $v_0^n = (-1)^n$.

Each plot shows the resulting distribution at time $t = .66$. This is an artificial experiment, since it amounts to specifying data on the outflow boundary of the interval $[-1, 0]$, but it highlights the completely predictable behavior of parasites. In Figs. 1.7a and 1.7b one sees waves of type $(\pi/h, 0)$ and $(0, 0)$ on the left and right, respectively. In Fig. 1.7c the waves have become of type $(0,\pi/k)$ and $(\pi/h,\pi/k)$, although to display the sawtooth behavior in t it would be necessary to show an additional plot for $t = .66 + k$. All of these waves travel at group speeds approximately ± 1. The remarkable x-symmetry in each plot is due to the ξ-symmetry about $\xi = \pi/2h$ of Fig. 1.1a, and the t-symmetry relating Figs. 1.7a and 1.7c is due to the corresponding ω-symmetry. These details are unimportant, for they would change with the difference scheme. What is important is that smooth behavior in either x or t is no guarantee of smooth behavior in the other variable, and that even extremely unphysical waves obey a group speed, which may have the wrong sign.

In problems involving parasitic waves the notion of phase speed is not just inadequate, but ill-defined. According to (1.2.1) the phase speed is $c = \omega/\xi$, but since ωk and ξh are only determined up to multiples of π, this formula does not give

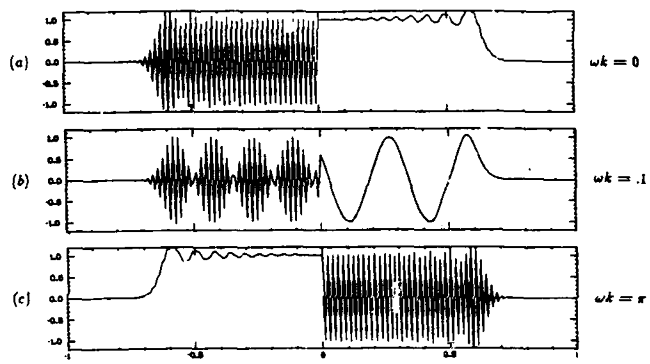


FIG. 1.7. Sawtoothed parasites generated by a forced oscillation $\sin \omega t$ at the middle of an interval, for various frequencies ω. In each case the forcing function was turned on at $t = 0$ and the result is plotted at $t = .66$. The model is LF for $u_t = -u_x$ with $h = 1/100$, $\lambda = .5$.

a unique value. The difficulty (regarding ωk) is that since the wave is only observable at discrete time intervals, it cannot be said whether a sine wave has moved left or right to get from one configuration to the next. But whatever phase speed one selects will fail to capture the basic fact of the speed at which the edge of the parasite moves. The group speed, by contrast, is well defined, because $d\omega/d\xi$ has the same value for all choices of ω and ξ.

The above examples have suggested that it is common for sawtoothed waves to propagate under nondissipative difference formulas in the wrong direction. It is convenient to devise a name for this property:

Defn. Let Q be a scalar difference formula. Suppose that whenever Q admits a solution $v_j^n = e^{i\omega t}$ with $\omega \in \mathbb{R}$ and group speed $C \in \mathbb{R}$, then it also admits the solution $v_j^n = (-1)^j e^{i\omega t}$, and this wave has group speed $C' \in \mathbb{R}$ satisfying $CC' \leq 0$, with $C' \neq 0$ if $C \neq 0$. Then Q is **x-reversing**. Likewise if the existence of a solution $v_j^n = e^{-i\xi x}$ with $\xi \in \mathbb{R}$ and group speed C implies the existence of a solution $v_j^n = (-1)^n e^{-i\xi x}$ with $CC' \leq 0$, with $C' \neq 0$ if $C \neq 0$, then Q is **t-reversing**. //

One may show readily for the schemes we have considered (see also App. A):

Theorem 1.5.1.

(i) LF and LF4 are both x-reversing and t-reversing,

(ii) BE and CN are x-reversing but not t-reversing,

(iii) LFd is t-reversing but not x-reversing,

(iv) LW is neither x-reversing nor t-reversing.

Proof. Let us prove *(ii)* for the scheme CN. Suppose $v_j^n = e^{i\omega t}$ satisfies CN with $\omega \in \mathbb{R}$. Then v has $\xi = 0$ by definition, so (1.1.18) implies $\tan \frac{\omega k}{2} = 0$, hence $\omega = 0$, and by (1.2.7) the solution has $C = -a \in \mathbb{R}$. The dispersion relation (1.1.18) implies that $v_j^n = (-1)^j$ is also a solution, with $\xi h = \pi$, and by (1.2.7) this solution has $C' = +a \in \mathbb{R}$, yielding $CC' = -a^2 < 0$. Therefore CN is x-reversing. On the other hand $v_j^n \equiv 1$ satisfies (1.1.18) but $v_j^n = (-1)^n$ (i.e. $\xi h = 0$, $\omega k = \pi$) does not, so CN is not t-reversing.

For the other assertions the proof is similar. ∎

Not every nondissipative difference formula is x-reversing. One way to see this is to observe that a centered spatial difference operator

$$a\frac{\partial}{\partial x} \approx \sum_{j=1}^{\ell} a_j \frac{(K^j - K^{-j})}{2jh}, \tag{1.5.1}$$

where K denotes the shift operator $Kv_j = v_{j+1}$, leads to a spatial factor

$$-\sum_{j=1}^{\ell} a_j \frac{\sin j\xi h}{jh}$$

in the dispersion relation. A difference formula based on this spatial discretization will have

$$C(\xi, 0) = -\sum_{j=1}^{\ell} a_j \cos j\xi h.$$

Consistency implies

$$C(0,0) = -\sum_{j=1}^{\ell} a_j = -a \tag{1.5.2}$$

but it implies nothing about the group velocity for a spatial sawtooth,

$$C(\pi/h, 0) = -\sum_{j=1}^{\ell} (-1)^j a_j. \tag{1.5.3}$$

Thus for exampl · formula

$$v_j^{n+1} - v_j^{n-1} = \frac{\lambda a}{3}(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda a}{3}(v_{j+2}^n - v_{j-2}^n)$$

has $a_1 = a/3$, $a_2 = a/3$, hence $C(\pi/h, 0) = -a/3 < 0$ as well as $C(0,0) = -a < 0$. But there will also be values ξ in $(0, \pi/h)$ with C of the opposite sign. Usually, for each frequency there will be as many wave numbers with $C < 0$ as with $C > 0$. Thus it is in the nature of nondissipative formulas to reverse some waves. In fact only a one-sided formula can fail to send some energy in the wrong direction, and such a formula is usually either unstable or dissipative. (However the Box scheme, listed in App. A, gives an example of a one-sided, nondissipative, not x-reversing formula.)

In practice, a nondissipative difference approximation to a first-order derivative will often be taken as the optimal formula for the given number of points. For the centered stencil of size $2\ell + 1$, this formula has order 2ℓ. (For example, LF and LF4 are based on x-difference approximations with $\ell = 1$ and $\ell = 2$, respectively.) In this important case, all formulas are reversing:

Theorem 1.5.2. *Let Q be a difference model of (1.1.1) whose spatial (resp. temporal) discretization consists of the optimal $2\ell + 1$-point centered difference approximation to $a\partial/\partial x$ (resp. $\partial/\partial t$). Then Q is x-reversing (resp. t-reversing).*

Proof. The optimal approximation in question can be given exactly ([Kr72], Remark p. 202): in the notation (1.5.1) one has

$$a_j = 2a(-1)^{j+1}\binom{\ell}{\ell - j}\bigg/\binom{\ell + j}{\ell}.$$

By (1.5.3) and the alternating signs of these coefficients, it is immediate that one has $C(\pi/h, 0)/a > 0$, and since $C(0,0)/a < 0$, the assertion is proved. ∎

As mentioned above, Chapter 4 will show that the stability of a difference model of an initial boundary value problem depends on whether the model can support waves with group velocity opposite to the physically correct direction. In practice, numerically unstable solutions often consist of sawtoothed waves under x- or t-reversing formulas, a fact that we will pursue in §4.4 and §4.5.

1.6 Wave propagation in several dimensions

Mathematically, linear wave propagation in several dimensions is much the same as in one, for the different space dimensions decouple. Nevertheless, the combination of these one-dimensional effects introduces geometrical phenomena that are surprising. In particular, difference schemes for isotropic equations are themselves anisotropic, and as a result imperfectly resolved waves travel not only at false speeds but in false directions. Such effects have received little treatment in the literature, but there are some previous studies, particularly by geophysicists [Al74,Ba80,Ma81,Wa80].* There is also a great deal known about wave propagation in crystal lattices, which is strongly analogous to propagation in finite difference grids, and there the same anisotropy phenomena appear. For references see for example [Au73,Bo54,Br53,Je37,So64].

In d dimensions, Fourier modes take the form

$$e^{i(\omega t-\xi\cdot x)}, \tag{1.6.1}$$

where ω is still a scalar frequency and ξ is now a **wave number vector** of dimension d. From (1.6.1) one may define the vector **phase velocity** c componentwise by

$$c_i = \omega\frac{\xi_i}{|\xi|^2} \qquad (1 \le i \le d). \tag{1.6.2}$$

The phase velocity points normal to the wave front, but has little physical significance. Once again, a stationary phase argument [Wh61,Wh74] can be used to show that energy travels at a **group velocity**, now given by

$$C = \nabla_\xi \omega, \tag{1.6.3}$$

*In geophysics one faces the inverse problem of inferring the properties of the earth from observations of sound propagation through it. On a global scale, the sound sources are earthquakes or nuclear explosions, and the goal is to understand the large-scale structure of the earth's surface or interior. On a scale of a few miles, the sound sources are dynamite explosions or other man-made impulses, and the goal is to detect inhomogeneities of sound speed that may gives clues to the location of oil or other resources. In these problems finite difference models are used extensively [Ba80,Cl76,Ma81]. The grids employed are often coarse relative to the wavelengths present, so numerical group velocity errors are potentially significant.

where ∇_ξ denotes the gradient (d-vector) with respect to ξ.

For simplicity, let us confine ourselves to two dimensions, and write (ξ,η) for ξ. Consider the (isotropic) ***second-order wave equation***

$$u_{tt} = u_{xx} + u_{yy}. \tag{1.6.4}$$

The dispersion relation for (1.6.4) is a system of concentric circles,

$$\omega^2 = \xi^2 + \eta^2, \tag{1.6.5}$$

which has two frequencies for each wave number because (1.6.4) is of second order. From (1.6.3) one obtains a group velocity

$$C = \pm\xi/|\xi|,$$

which asserts that energy travels normal to the wave front at speed 1. As a typical finite difference model of (1.6.4), suppose we define a rectilinear grid with step size h in both x and y, and consider **second-order leap frog** (LF^2):

$$v_{ij}^{n+1} - 2v_{ij}^n + v_{ij}^{n-1} = \lambda^2[v_{i+1,j}^n + v_{i-1,j}^n + v_{i,j+1}^n + v_{i,j-1}^n - 4v_{ij}^n]. \tag{1.6.6}$$

(The restriction of this formula to one dimension is included in the summary of Appendix A.) Easy trigonometric manipulations then yield the numerical dispersion relation (cf. [Al74], eq. (A2))

$$\sin^2\frac{\omega k}{2} = \lambda^2\left[\sin^2\frac{\xi h}{2} + \sin^2\frac{\eta h}{2}\right]. \tag{1.6.7}$$

From a contour plot of (1.6.7), one can see the errors in group velocity that LF^2 will give rise to (cf. [Au73], [Je37, chap. 15]). Fig. 1.8 shows curves of constant ω in ξ-space for $\omega h = \pi/8,\ldots,11\pi/8$. For simplicity λ has been taken here equal to 0, so that LF^2 is reduced to a semi-discrete or "method of lines" approximation. The full domain portrayed is $(\xi,\eta) \in [-\pi/h,\pi/h]^2$ (in crystal terminology, the ***first Brillouin zone***); any other wave number vector is an alias of a vector in this region. The figure shows that as ω increases, the curve of corresponding ξ vectors becomes less like a circle and more like a diamond. Now (1.6.3) implies that the group velocity for any wave number ξ points in the direction of the normal to the line of constant ω through ξ. By contrast the phase velocity, since it is normal to the wave front, lies along the ray from the origin through ξ, and so would the ideal group velocity for

(1.6.4). Thus Fig. 1.8 indicates that poorly resolved wave packets will travel more along a diagonal under LF^2 than they ought to. The figure also shows an increasing separation between curves of constant ω as ω increases. By (1.6.3) this indicates that poorly resolved packets will travel too slowly, as in the one-dimensional case, and evidently this effect will be more pronounced at 0° or 90° than at 45°.

Applying (1.6.3) to (1.6.7) recapitulates these phenomena algebraically (cf. [AJ74]). One obtains the group velocity components

$$C_x = \frac{\lambda \sin \xi h}{\sin \omega k}, \qquad C_y = \frac{\lambda \sin \eta h}{\sin \omega k}. \tag{1.6.8}$$

Therefore the group propagation angle (from the x axis) and speed for the wave number vector (ξ, η) are

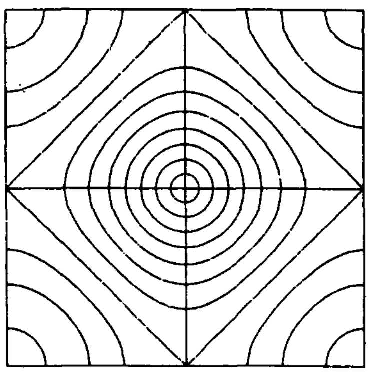


FIG. 1.8. Dispersion plot for the two-dimensional Leap Frog model of $u_{tt} = u_{xx} + u_{yy}$ in the limit $\lambda \to 0$. The region shown is the domain $[-\pi/h, \pi/h]^2$ of the $\xi = (\xi, \eta)$ plane. The concentric curves plotted are lines of constant ω for $\omega h = \pi/8, 2\pi/8, \ldots, 11\pi/8$. The normal to the curve passing through a point ξ is the direction of the corresponding group velocity.

$$\Theta = \tan^{-1}\left(\frac{\sin \eta h}{\sin \xi h}\right), \tag{1.6.9}$$

$$|C| = \frac{\lambda\sqrt{\sin^2 \xi h + \sin^2 \eta h}}{\sin \omega k}. \tag{1.6.10}$$

For infinitesimal ξh these expressions reduce to the isotropic and nondispersive formulas

$$\Theta = \tan^{-1}\frac{\eta}{\xi}, \qquad |C| = 1,$$

but for finite ξh they confirm that there is anisotropy and dispersion. Let θ denote the angle from the x axis of the normal to a given plane wave. Then to second order one has

$$|C| \approx 1 - \frac{(|\xi|h)^2}{t}\left[\frac{3 + \cos 4\theta}{4} - \lambda^2\right], \tag{1.6.11}$$

$$\Theta \approx \theta + \frac{(|\xi|h)^2}{24}\sin 4\theta. \tag{1.6.12}$$

Eq. (1.6.12) shows again that waves will travel more slowly than the correct speed 1, lagging twice as much (for small λ) at $\theta \equiv 0°$ (mod 90°) as at $\theta \equiv 45°$ (mod 90°). Eq. (1.6.13) confirms that waves with $\theta \equiv 0°$ (mod 45°) will propagate perpendicularly to the wave front (a fact obvious from the symmetries of the grid), but that all other waves will propagate obliquely, preferring diagonals to horizontals and verticals. The details would change if the x and y mesh spacings were not equal.

DEMONSTRATION 1.7. Fig. 1.9 confirms these predictions experimentally. Here a Gaussian wave packet

$$u(x, 0) = \sin(x \cdot \xi)e^{-30|x|^2}$$

with $\theta = 22.5°$ and $|\xi|h = 1.6$ has been set up at $t = 0$. The experiment takes $h = .01$, $\lambda = .4$, scheme $= \mathrm{LF}^2$. Superimposed on the same plot is the packet at the later time $t = 1.4$. Ideally it should have traveled a distance 1.4 at an angle 22.5°. In fact, it has closely matched the predictions of (1.6.9) and (1.6.10): $\Theta = 30.0°$, $|C| = .81$.

In realistic problems, coefficients will usually vary in space. Following a standard theory of **ray tracing** in inhomogeneous anisotropic media [Li78], it is possible to work out in detail what kind of errors discretization will introduce. Now one has a space-dependent dispersion relation

$$\omega = \omega(x, \xi)$$

and the group velocity formula (1.6.3) becomes half of a system of equations in Hamiltonian form,

$$\frac{dx}{dt} = \nabla_\xi \omega, \qquad \frac{d\xi}{dt} = -\nabla_x \omega. \tag{1.6.13}$$

In the special case of a **stratified medium**, in which the spatial dependence involves one dimension only, one can simplify this system by replacing the second equation by an algebraic formula $\xi = \xi(x)$ derived from the numerical dispersion relation, and this is a numerical form of **Snell's Law**. For an example, see [Tr82]. Some further remarks on Snell's Law are given at the end of §3.6.

One might go further, and study wave propagation in nonlinear models by means of the fairly well developed theory of nonlinear wave propagation in dispersive media [Wh74]. However, we will not pursue this here.

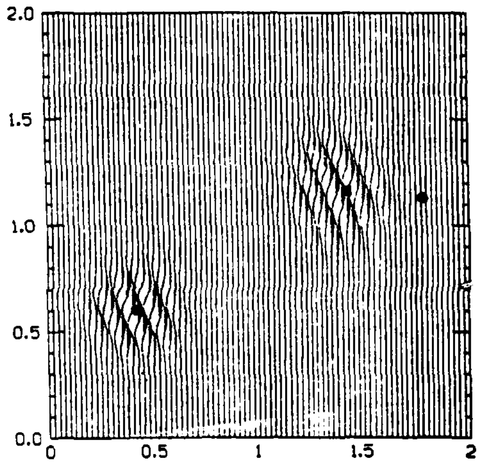


FIG. 1.9. Propagation of a two-dimensional wave packet with $|\xi|h = 1.6$, $\theta = 22.5°$. The model is the Leap Frog scheme for $u_{tt} = u_{xx} + u_{yy}$ with $h = 1/100$, $\lambda = .4$. The packet is shown at both $t = 0$ (lower left) and $t = 1.4$ (upper right). Dots mark the ideal starting and ending positions, and the square the position predicted by (1.6.8)–(1.6.10).

2. LEFTGOING AND RIGHTGOING SIGNALS

2.1 The general scalar difference formula

The purpose of this chapter is to make the results presented so far more general and more rigorous. The key to this is an algebraic study of the dispersion relation for an arbitrary scalar difference formula Q—two-level or multilevel, explicit or implicit. For a complete analysis one must permit ω and ξ to be complex, and one must examine the defective solutions that occur when ω or ξ has multiplicity greater than 1. In this first section we will define Q and describe the solutions it admits with regular behavior in x and t (Thms. 2.1.1,2.1.2). Section 2.2 details the relationships of wave number and frequency to *x-dissipativity*, *t-dissipativity*, and *Cauchy stability*. Section 2.3 then sets forth our most important foundational material for Chapters 3-6. First, Thm. 2.3.1 proves that if Q is Cauchy stable, then the dispersion relation is analytic about any point with $\xi, \omega \in \mathbb{R}$, and there exists a real group velocity. Second, Thm. 2.3.2 describes the connection between wavelike modes, with $\omega, \xi \in \mathbb{R}$, and evanescent modes, with ω or ξ complex. These results form the basis of definitions of ***rightgoing*** and ***leftgoing***, ***strictly rightgoing*** and ***strictly leftgoing***, and ***stationary*** solutions to Q, which will be central to our later work on boundaries, interfaces, and stability. Section 2.4 then goes on to apply these results to the class of ***three-point linear multistep formulas***, and Section 2.5 extends them to diagonalizable vector difference models.

We begin by introducing space and time shift operators:

Defn. The shift operators K and Z are defined by*

*To avoid abuse of notation, we would have to be consistent as to whether v is a doubly indexed sequence, a time sequence of space sequences $(v_j)^n$, or a space sequence of time sequences $(v^n)_j$. Unfortunately any such fixed choice is too cumbersome to be practical, and we will apply K freely to any object that has a spatial index, and Z to any object with a time index.

$$K v_j^n = v_{j+1}^n, \qquad Z v_j^n = v_j^{n+1}. \quad //$$

Let us define complex numbers κ, z by

$$\kappa = e^{-i\xi h}, \qquad z = e^{i\omega k}. \tag{2.1.1}$$

Then the Fourier mode (1.1.4) takes the form

$$v_j^n = e^{i(\omega t - \xi x)} = z^n \kappa^j \qquad (x = jh,\ t = nk), \tag{2.1.2}$$

and it is an eigenfunction of K and Z with eigenvalues κ, z. The case in which ξ or ω is real corresponds to the situation $|\kappa| = 1$ or $|z| = 1$, respectively. In this dissertation we will use ξ, ω or κ, z, or both, according to convenience. This use of κ and z follows the stability work of Kreiss and colleagues [Gu72,etc.], and we have introduced K and Z by analogy. The remaining ideas of this section are also heavily influenced by those of [Gu72].

A general $s+2$-level finite difference model Q of (1.1.1) with constant coefficients can be written

$$Q_{-1} v^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma v^{n-\sigma}, \tag{2.1.3}$$

where each Q_σ is a spatial difference operator,

$$Q_\sigma = \sum_{j=-\ell}^{r} a_{j\sigma} K^j \qquad (-1 \le \sigma \le s). \tag{2.1.4}$$

We assume that Q_{-1} has a bounded inverse in ℓ_2. If $Q_{-1} = 1$, (2.1.3) is **explicit**; otherwise it is **implicit**. We assume that $\lambda = k/h$ is fixed and that the coefficients $a_{j\sigma}$ depend on λ, but not on h and k independently. The integers $\ell \ge 0$ and $r \ge 0$ define how far left and right the stencil extends.

Carrying the shift operator notation further, we can write Q in the form

$$P(K, Z)v = \left[\sum_{j=-\ell}^{r} \sum_{\sigma=-1}^{s} a_{j\sigma} K^{j+\ell} Z^{s-\sigma}\right] v = 0, \tag{2.1.5}$$

where P is a bivariate polynomial of degree $\ell + r$ with respect to K and degree $s+1$ with respect to Z. The dispersion relation for Q is then simply

$$P(\kappa, z) = 0. \tag{2.1.6}$$

In this notation LF takes the form

$$[K(Z^2-1)-\lambda a Z(K^2-1)]v=0,$$

or equivalently,

$$[(Z-Z^{-1})-\lambda a(K-K^{-1})]v=0, \tag{2.1.7}$$

and its dispersion relation (1.1.7) becomes

$$z-\frac{1}{z}=\lambda a(\kappa-\frac{1}{\kappa}). \tag{2.1.8}$$

Similarly LW has the shift operator form

$$\left[ZK-K-\frac{\lambda a}{2}(K^2-1)-\frac{(\lambda a)^2}{2}(K^2-2K+1)\right]v=0,$$

that is,

$$\left[Z-1-\frac{\lambda a}{2}(K-K^{-1})-\frac{(\lambda a)^2}{2}(K-2+K^{-1})\right]v=0. \tag{2.1.9}$$

In these instances the space and time parts of the difference formula are independent. We define in general

Defn. The formula Q is **separable** if it can be written in the form

$$[f(Z)-g(K)]v=0, \tag{2.1.10}$$

where f and g are rational functions. //

LF, LW, and many other difference formulas used in practice are separable. For example, CN can be written

$$\left[\frac{Z-1}{Z+1}-\frac{\lambda a}{4}(K-1/K)\right]v=0, \tag{2.1.11}$$

and LF4 has the form

$$\left[Z-1/Z-\frac{4\lambda a}{3}(K-1/K)+\frac{\lambda a}{6}(K^2-1/K^2)\right]v=0. \tag{2.1.12}$$

Any difference formula based on the *method of lines*, in which the x discretization is carried out before the t discretization, will also be separable. An example of a nonseparable scheme is LFd (§1.1), which has the shift operator form

$$\left[Z-1/Z-\lambda a(K-1/K)-\frac{\epsilon}{16Z}(K-1)^2(1-1/K)^2\right]v=0. \tag{2.1.13}$$

Separable schemes have the property that their group velocities factor into a product

$$C(\omega,\xi)=C_1(\omega)C_2(\xi). \tag{2.1.14}$$

We have observed this for particular examples in (1.2.5), (1.2.7), and (1.2.8). The reason for the factorization in general is that if Q is separable, its dispersion relation can be written

$$f(e^{i\omega k})=g(e^{-i\xi h}).$$

Differentiation gives

$$ike^{i\omega k}f'(e^{i\omega k})d\omega=-ihe^{-i\xi h}g'(e^{-i\xi h})d\xi,$$

and hence by (1.2.3),

$$C=\frac{d\omega}{d\xi}=-\frac{1}{\lambda}\Big(e^{i\omega k}f'(e^{i\omega k})\Big)^{-1}\Big(e^{-i\xi h}g'(e^{-i\xi h})\Big).$$

We will be extensively concerned with the relation between κ and z imposed by the dispersion relation (2.1.6). To begin with, suppose that κ is fixed. We ask the question: what solutions of the form

$$v_j^n=\kappa^j\psi_n, \tag{2.1.15}$$

where $\{\psi_n\}$ is a sequence in n, does Q support? By (2.1.5), (2.1.15) is a solution of $Qv=0$ if and only if

$$P(\kappa,Z)\psi_n=0. \tag{2.1.16}$$

This is an ordinary difference equation for ψ_n, and the solutions to such equations are well known:

Theorem 2.1.1. *Let $\kappa\in\mathbb{C}$ be arbitrary, and assume that the polynomial*

$$P_\kappa(z)=P(\kappa,z)$$

is of exact degree $s+1$, i.e. the coefficient of the z^{s+1} term is nonzero. Let $\{z_i\}_{1\le i\le\mu}$ denote its distinct roots, with z_i of multiplicity ν_i, hence $\sum_{i=1}^{\mu}\nu_i=s+1$. Then the $s+1$ sequences

$$\psi_n=z_i^n n^\delta \qquad 1\le i\le\mu,\quad 0\le\delta\le\nu_i-1 \tag{2.1.17}$$

are linearly independent solutions of (2.1.16), and they span the linear space of all such solutions.

Proof. [Or72], §4.2. ▮

Remark. By assumption Q_{-1} is invertible in ℓ_2, from which it follows that for $|\kappa| = 1$ (and hence, by continuity, for $|\kappa|$ sufficiently close to 1), the assumption of exact degree $s+1$ must hold.

Now let us switch the roles of κ and z, and suppose z is fixed. Corresponding to (2.1.15), we may ask, what solutions

$$v_j^n = z^n \phi_j, \tag{2.1.18}$$

where $\{\phi_j\}$ is a sequence in j, does Q support? For this one has corresponding to (2.1.16) the equation

$$P(K, z)\phi_j = 0, \tag{2.1.19}$$

which is called the **resolvent equation** for Q ([Gu72], eq. (4.1)). Again we have an ordinary difference equation whose solution can be characterized completely:

Theorem 2.1.2. *Let $z \in \mathbb{C}$ be arbitrary, and assume that the polynomial*

$$P_z(\kappa) = P(\kappa, z)$$

is of exact degree $\ell + r$. Let $\{\kappa_i\}_{1 \le i \le \mu}$ denote its distinct roots, with κ_i of multiplicity ν_i, hence $\sum_{i=1}^{\mu} \nu_i = \ell + r$. Then the $\ell + r$ sequences

$$\phi_j = \kappa_i^j j^\delta \qquad 1 \le i \le \mu, \quad 0 \le \delta \le \nu_i - 1 \tag{2.1.20}$$

are linearly independent solutions of (2.1.19), and they span the linear space of all such solutions.

Proof. Same as for Thm. 2.1.1. ▮

This theorem, which we will use more often than Thm. 2.1.1, provides a complete breakdown of all solutions with regular time behavior that Q can support. In later sections the analysis usually comes down to determining which combinations of these solutions are permitted by particular choices of boundary or interface conditions.

2.2 Cauchy stability and dissipativity

We will be concerned only with difference formulas that are ℓ_2-stable in the absence of boundaries or interfaces. The following definition is the same as the definition of L_2 stability in §1.4, except that L_2 is replaced by ℓ_2 and we now cover the case of multilevel formulas.

Defn. Q is **Cauchy stable** if for each $T > 0$, there exists a constant $C_T > 0$ such that

$$\|v^n\|_2 \le C_T \sum_{\sigma=0}^{s} \|v^\sigma\|_2$$

for all n and k satisfying $nk \le T$, where $\|\cdot\|_2$ denotes the norm defined by

$$\|\phi\|_2^2 = h \sum_{j=-\infty}^{\infty} |\phi_j|^2. \quad // \tag{2.2.1}$$

The results of the last section lead to necessary conditions for Cauchy stability. Here and in later sections, when we speak of connections between κ and z, it should be understood that we are concerned only with pairs (κ, z) that satisfy the dispersion relation (2.1.6).

Defn. The model Q satisfies the **von Neumann condition** if $|\kappa| = 1$ implies $|z| \le 1$. //

Theorem 2.2.1. *A necessary condition for Cauchy stability is that Q satisfies the von Neumann condition. A further necessary condition is that $|\kappa| = |z| = 1$ implies that z is simple.**

Proof. If the von Neumann condition does not hold, then by Thm. 2.1.1, Q admits a solution

$$v_j^n = \kappa^j z^n, \qquad n \ge 0$$

with $|\kappa| = 1$ and $|z| > 1$. If the simple root condition does not hold, the same theorem shows that Q admits a solution

$$v_j^n = n\kappa^j z^n, \qquad n \ge 0$$

with $|\kappa| = |z| = 1$. It follows that in either case the nth powers of the amplification matrices corresponding to the Fourier mode ξ with $e^{-i\xi h} = \kappa$ grow unboundedly as

*In fact in the present constant-coefficient situation, the conditions given are also sufficient for stability. But we will not need this result.

$n \to \infty$ for fixed k. Therefore they also grow unboundedly as $k \to 0$ for fixed T. Since such amplification matrices are continuous functions of ξ, Cauchy instability follows by Fourier analysis (see §5.4 of [Ri67]). ∎

The definition of dissipativity is a further strengthening of the von Neumann condition:

Defn. Q is **dissipative** or **x-dissipative** if it satisfies the von Neumann condition, and moreover, $|\kappa| = 1$, $\kappa \neq 1$ implies $|z| < 1$, or equivalently, $|\kappa| = |z| = 1$ implies $\kappa = 1$. It is **strictly nondissipative** or **unitary** if $|\kappa| = 1$ implies $|z| = 1$. //

Note that strict nondissipativity is a stronger condition than the negative of dissipativity. Most formulas are one or the other, but an example of one that falls between is BE (§1.1). For $|\kappa| = 1$, $\kappa \neq \pm 1$, BE has $|z| < 1$, but the mode $\kappa = -1$, $z = 1$ keeps it from being x-dissipative.

In practice, what one often needs is a slightly stronger property:

Defn. Q is **totally dissipative** if it satisfies the von Neumann condition and moreover, $|\kappa| = |z| = 1$ implies $\kappa = z = 1$. //

For two-level schemes ($s = 0$), we will show in a moment that x-dissipativity and total dissipativity are equivalent. An example suffices to show that for multilevel schemes the situation is different: LFd (§1.1) is x-dissipative, but it admits the mode $\kappa = 1$, $z = -1$, so it is not totally dissipative. The fact that x-dissipativity does not ensure total dissipativity for multilevel schemes causes occasional confusion and error in papers on finite difference methods, which is why we choose to add the prefix x.

In analogy, one might define a ***t*-dissipative** formula to be one for which $|\kappa| = |z| = 1$ implies $z = 1$. For generality in later applications (see especially §6.2), we choose to make the definition slightly weaker—the minimum necessary so that x- and t- dissipativity together imply total dissipativity. The following definition is closely related to condition (3.7) in the paper [Go81] by Goldberg and Tadmor, and to the notion of ***tangential dissipativity*** introduced by Coughran in [Co80].

Defn. [Go81], eq. (3.7)). Q is **t-dissipative** if $\kappa = 1, |z| = 1$ implies $z = 1$. //

Thus, for example, BE is t-dissipative but not x-dissipative.

Theorem 2.2.2. *Q is totally dissipative if and only if it is both x-dissipative*

and t-dissipative.

Proof. Both total dissipativity and x-dissipativity require the von Neumann condition, so that part of the equivalence holds. What remains is to show that $|z| = |\kappa| = 1 \Rightarrow z = \kappa = 1$ is equivalent to $|z| = |\kappa| = 1 \Rightarrow \kappa = 1$ plus $\kappa = 1, |z| = 1 \Rightarrow z = 1$. This is immediate. ∎

The example of LFd showed that x-dissipativity does not imply t-dissipativity. However, for two-level schemes one has

Theorem 2.2.3. *Any consistent two-level formula Q is t-dissipative. Any consistent two-level x-dissipative formula Q is totally dissipative.*

Proof. By consistency, Q must have a solution $\kappa = z = 1$, and if it is a two-level formula, then by Thm. 2.1.1 there can only be a single z for each κ, so this is the only solution with $\kappa = 1$. Therefore the condition of t-dissipativity holds trivially. If Q is also x-dissipative, then it is totally dissipative by Thm. 2.2.2. ∎

One readily sees that dissipativity precludes the possibility that a scheme is reversing:

Theorem 2.2.4. *If Q is consistent and x-dissipative, it cannot be x-reversing. If Q is consistent and t-dissipative, it cannot be t-reversing.*

Proof. If Q is consistent, then $\kappa = z = 1$ is a solution with $C = -a \in \mathbb{R}$. For Q to be x-reversing, it must therefore admit the solution $\kappa = -1, z = 1$. This contradicts the definition of an x-dissipative formula. Similarly for the t case. ∎

For a scalar difference model with constant coefficients, dissipativity almost completely determines the behavior and stability of solutions to the Cauchy problem in the ℓ_2 or L_2 norms. In the two-level case, its influence is complete. Each Fourier component κ will lose L_2 energy at the rate $|z(\kappa)|^n$, and by Parseval's formula, the overall solution will decay according to the combination of these effects. One might say that dissipation acts on individual wave numbers independently, and the L_2 norm measures them independently. For problems with variable coefficients, two important theorems of Kreiss [Ri67,§6] show that dissipativity still goes a long way towards ensuring L_2-stability.

Dispersion, on the other hand, has to do with the interaction of wave numbers, and the results of §1.4 show that this interaction must be taken into account for stability in L_p norms other than L_2. We will see that the same is true, even the in L_2

norm. for problems containing boundaries or interfaces.

2.3 Leftgoing and rightgoing solutions

We now have the material in place to return to group velocity and give it a fuller explanation. First, the following theorem establishes that "group velocity always makes sense"—for any wavelike mode, the derivative (1.2.3) exists and is real.

Theorem 2.3.1. *Let Q be a Cauchy stable scalar difference formula with constant coefficients, as described in §2.1. Suppose that Q admits a solution*

$$v_j^n = z_0^n \kappa_0^j = e^{i(\omega_0 t - \xi_0 x)} \qquad (x = jh,\ t = nk) \tag{2.3.1}$$

with $|z_0| = |\kappa_0| = 1$, i.e. $\omega_0, \xi_0 \in \mathbb{R}$. Then

(i) In a neighborhood of (κ_0, z_0), z is a single-valued analytic function of κ.

(ii) The group velocity derivative $C = d\omega/d\xi$ exists at (κ_0, z_0), and is real.

(iii) $C(\kappa_0, z_0) = 0$ if and only if κ_0 is multiple (i.e. a multiple root of the polynomial $P_{z_0}(\kappa) = P(\kappa, z_0)$ of §2.1).

Proof. If Q admits the solution (2.3.1), then $P(\kappa_0, z_0) = 0$, where P is the bivariate polynomial defined in (2.1.5). By the remark following Thm. 2.1.1, the univariate polynomial $P_\kappa(z) = P(\kappa, z)$ has exact degree $s+1$ for all κ in a neighborhood of $\kappa = \kappa_0$, and by the definition of P, its coefficients are analytic functions of κ (in fact polynomials). Moreover since Q is Cauchy stable, Thm. 2.2.1 implies that z_0 is a simple root of $P_{\kappa_0}(z)$. From these facts it follows by the implicit function theorem that in a neighborhood of (κ_0, z_0), the equation $P(\kappa, z) = 0$ determines a unique analytic function $z(\kappa)$, satisfying

$$(z - z_0) = A(\kappa - \kappa_0)^\nu + O((\kappa - \kappa_0)^{\nu+1}) \qquad A \neq 0, \tag{2.3.2}$$

for some $A \in \mathbb{C}$, where $\nu \geq 1$ is the multiplicity of κ_0 as a root of $P_{z_0}(\kappa) = P(\kappa, z_0)$. This proves *(i)*.

By differentiating (2.1.1), one obtains the formulas

$$d\kappa = -ih\kappa\, d\xi, \qquad dz = ikz\, d\omega. \tag{2.3.3}$$

Since we have shown that $dz/d\kappa$ exists at (κ_0, z_0), it follows that $C(\kappa_0, z_0)$ exists and is given by the formula

$$C(\kappa_0, z_0) = \left.\frac{d\omega}{d\xi}\right|_{\xi_0, \omega_0} = -\frac{1}{\lambda}\frac{dz}{d\kappa}\frac{\kappa}{z}\bigg|_{\kappa_0, z_0}. \tag{2.3.4}$$

By (2.3.2), $C(\kappa_0, z_0) = 0$ if and only if $\nu \geq 2$, which proves *(iii)*.

Assume on the other hand $\nu = 1$, so that $z'(\kappa_0) \neq 0$, and (2.3.2) and (2.3.4) give

$$C(\kappa_0, z_0) = -\frac{1}{\lambda} A \frac{\kappa_0}{z_0}. \tag{2.3.5}$$

Figure 2.1 indicates the situation—the function $z(\kappa)$ maps a neighborhood of κ_0 conformally onto a neighborhood of z_0.

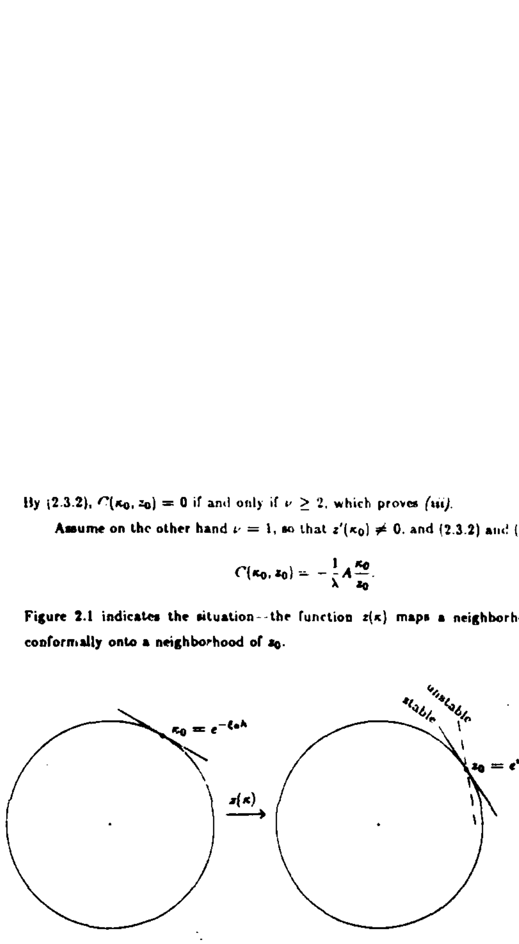


FIG. 2.1

For Cauchy stability the von Neumann condition must be satisfied, which means that $z(\kappa)$ must map $|\kappa| = 1$ into $|z| \leq 1$. Obviously, this can only happen if $z(\omega)$ maps the tangent to $|\kappa| = 1$ at κ_0 onto a curve that is tangent to $|z| = 1$ at z_0, as indicated in the Figure. This tangency condition is the same as the condition that the right hand side of (2.3.4) is real. This completes the proof of (ii). ∎

The significance of this theorem is that it applies to *all* wavelike solutions, including those involving defective roots κ and those admitted by formulas that are x- or t-dissipative. For example, BE admits the wave $(-1)^j$ and LFd admits the wave $(-1)^n$, as mentioned in §1.4, but most solutions with $|\kappa| = 1$ under these formulas have $|z| < 1$. Thm. 2.3.1 shows that nevertheless, these waves have well defined group velocities. (For another example, see the Lax-Friedrichs scheme listed in App. A.) Though we will not give any details until Appendix B, the stationary phase argument of §1.2 or other related arguments confirm that these group velocities correctly describe the propagation of energy in these modes.

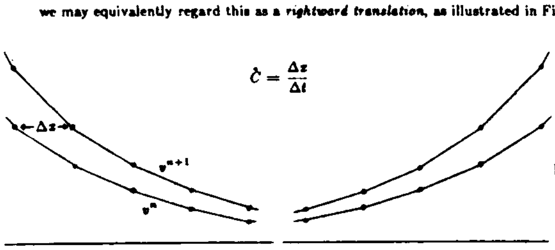
What Thm. 2.3.1 does not do is assign a group speed to signals with $|z| \neq 1$ or $|\kappa| \neq 1$. We will now show that for $|z| \geq 1$ and $|\kappa| \neq 1$, there is a speed of translation

$\check{C}$ naturally associated with a signal $z^n\kappa^j$, and this speed approaches C in the limit $|z| \downarrow 1$, $|\kappa| \uparrow 1$.

Let Q again be a Cauchy stable formula as in §2.1, and suppose it admits a solution

$$v_j^n = z^n \phi_j = z^n \kappa^j j^\delta \tag{2.3.6}$$

as described in Thm. 2.1 2, with $|z| > 1$. By Thm. 2.2.1, we must have either $|\kappa| < 1$ or $|\kappa| > 1$. Let us suppose $|\kappa| < 1$, and assume first $\delta = 0$. Now from one step to the next, the envelope $|v_j|$ increases by the fixed factor $|z|$ at all points j. However, we may equivalently regard this as a *rightward translation*, as illustrated in Fig. 2.2a.



(a) (b)

FIG. 2.2

In order for $|v|$ to increase by the factor $|z|$, this translation must cover a distance Δx satisfying

$$|\kappa|^{-\Delta x/h} = |z|,$$

that is,

$$\Delta x = -h \frac{\log|z|}{\log|\kappa|}.$$

Since the time step has length $k = \lambda h$, this amounts to rightward motion at a speed

$$\check{C} = \frac{\Delta x}{\Delta t} = -\frac{1}{\lambda}\frac{\log|z|}{\log|\kappa|}. \tag{2.3.7}$$

For $|\kappa| > 1$, illustrated in Fig. 2.1b, the situation is similar and we have leftward motion at a speed given by the same formula. Eq. (2.3.7) also applies to signals with $|z| = 1$ and $|\kappa| \neq 1$, where it gives the result $\check{C} = 0$.

In the defective situation $\delta \geq 1$, we can still view the evolution with time as a rightward or leftward motion, now coupled with a lower-order change of shape. One

way to make this motion quantitative would be to measure the increase in the total ℓ_2 energy to the right of a fixed point j (or to the left, for a leftgoing signal) from one step to the next (see §3.3). However, we will not pursue this.

Here is the result on $\check{C} \to C$ and related matters:

Theorem 2.3.2. *Let Q be a Cauchy stable difference formula as in Thm. 2.3.1, and suppose again that Q admits a solution*

$$v_j^n = z_0^n \kappa_0^j \tag{2.3.8}$$

with $|\kappa_0| = |z_0| = 1$. Let κ_0 have multiplicity $\nu \geq 1$. Let Ω denote the intersection of $\{z \in \mathbb{C} : |z| > 1\}$ with a neighborhood of $z = z_0$ chosen small enough so that for z in that neighborhood, the map $z(\kappa)$ of Thm. 2.3.1 defines ν continuous functions $\{\kappa_i(z)\}_{1\leq i\leq \nu}$ with $\kappa_i(z) \to \kappa_0$ as $z \to z_0$.

(i) For each i, either $|\kappa_i(z)| < 1$ $\forall z \in \Omega$ or $|\kappa_i(z)| > 1$ $\forall z \in \Omega$. Let ν_r denote the number of κ's in the former category and ν_ℓ the number in the latter (hence $\nu = \nu_\ell + \nu_r$). Then if ν is even, $\nu_\ell = \nu_r = \nu/2$; if ν is odd, either $\nu_\ell = (\nu+1)/2$ and $\nu_r = (\nu-1)/2$, or the reverse.

(ii) Let $\check{C}_i(z)$ denote the translation speed (2.3.7) for the signal $z^n\kappa_i^j(z)$. Then

$$\lim_{\substack{z\to z_0\\ z\in\Omega}} \check{C}_i(z) = C(\kappa_0, z_0) \tag{2.3.9}$$

for each i.

*(iii) (**Perturbation test**) If $C(\kappa_0, z_0) \neq 0$ (so that by Thm. 2.3.1, $\nu = 1$, and we can write $\kappa(z)$ for $\kappa_1(z)$), then $C(\kappa_0, z_0) > 0$ iff $|\kappa(z)| < 1$ for $z \in \Omega$, and $C(\kappa_0, z_0) < 0$ iff $|\kappa(z)| > 1$ for $z \in \Omega$. That is, $C(\kappa_0, z_0)$ is negative if $\nu_\ell = 1$ and positive if $\nu_r = 1$.*

Proof. The result $|\kappa_i(z)| \neq 1$ for $z \in \Omega$ follows from the von Neumann condition together with the fact that $\kappa_i(z)$ is a continuous function of z. The rest of claim (i) is implied by (2.3.2) (cf. Thm. 9.2 of [Gu72]).

The proof of (ii) requires only an algebraic verification. If κ_0 is multiple, then (2.3.2) and (2.3.7) imply $\lim_{z\to z_0} \check{C}_i(z) = 0$, the correct value. If κ_0 is simple, then by (2.3.5) and (2.3.7), what needs to be shown amounts to

$$\lim_{z\to z_0} \frac{\log|z|}{\log|\kappa|} = A\frac{\kappa_0}{z_0}, \tag{2.3.10}$$

where A is the constant of (2.3.2). For $\kappa \in \kappa(\Omega)$, let us write

$$\kappa = \kappa_0(1 + \epsilon e^{i\phi})$$

with $\epsilon > 0$. Then by (2.3.2),

$$z = z_0 + A\kappa_0\epsilon e^{i\phi} + O(\epsilon^2) = z_0(1 + A\frac{\kappa_0}{z_0}\epsilon e^{i\phi}) + O(\epsilon^2).$$

These two formulas imply

$$\log|\kappa| = \epsilon\cos\phi + O(\epsilon^2)$$

and, since $A\kappa_0/z_0$ is known to be real by (2.3.5),

$$\log|z| = A\frac{\kappa_0}{z_0}\epsilon\cos\phi + O(\epsilon^2).$$

By taking the ratio of these equations, one obtains (2.3.10), and this proves (ii).

Claim (iii) is a corollary of (ii), using (2.3.7). ∎

The observation $\check{C} \to C$ amounts to our third explanation of group velocity, to supplement those presented in §1.2 (beating of two waves; stationary phase). The idea is simple: since a wave $e^{i(\omega t - \xi x)}$ has uniform envelope 1, one cannot see how fast the envelope is moving; as soon as ξ is made slightly complex, however, the envelope acquires shape and its motion becomes apparent. The perturbation test specializes this to the statement that if all one cares about is the direction of motion, then all one must check is whether $|\kappa| < 1$ or $|\kappa| > 1$ for $|z| > 1$.

Our goal in this section has been to set up definitions of *leftgoing* and *rightgoing* signals, which will be of critical importance. Here they are.

Defn. Let Q admit a solution v of the form (2.3.6) with $|z| \geq 1$ and $\delta \leq \max\{\nu_\ell, \nu_r\}$ (defined in Thm. 2.3.2).

(i) If $|z| > 1$ and $|\kappa| < 1$ (resp. $|\kappa| > 1$), or if $|z| = |\kappa| = 1$ and $C(\kappa, z) > 0$ (resp. $C(\kappa, z) < 0$), then v is **strictly rightgoing** (resp. **strictly leftgoing**).

(ii) If v is strictly rightgoing (resp. strictly leftgoing), or if $|z| = 1$ and $|\kappa| < 1$ (resp. $|\kappa| > 1$), or if $|z| = |\kappa| = 1$ and $C(z,\kappa) = 0$ and $\delta \leq \nu_r$ (resp. $\delta \leq \nu_\ell$), then v is **rightgoing** (resp. **leftgoing**).

(iii) If v is both rightgoing and leftgoing, it is **stationary**. (That is, v is stationary if $|\kappa| = |z| = 1$, $C(\kappa, z) = 0$, and $\delta \leq \min\{\nu_\ell, \nu_r\}$.) //

These definitions divide the set of solutions (2.3.6) with $\delta \leq (\nu + 1)/2$ into nine classes, ranging from the strictly leftgoing mode of Fig. 2.2b to the strictly rightgoing mode of Fig. 2.2a. Table 2.1 summarizes this classification. We will see in §4 and §5

that positions (5) through (9) in the table are associated with increasing degrees of instability for initial boundary value problems.

The distinctions between positions (4), (5) and (6) of Table 2.1, based on Thm. 2.3.2, are perhaps the most difficult to grasp. Table 2.2 clarifies the situation by illustrating the connection of (4), (5), and (6) to the behavior of the dispersion relation in the vicinity of a point with $\xi, \omega \in \mathbb{R}$. The figure makes it clear why in the case of ν odd, the numbers of leftgoing and rightgoing modes are unequal.

TABLE 2.1

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\|z\| > 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| = 1$	$\|z\| > 1$
$\|\kappa\| > 1$	$\|\kappa\| = 1$	$\|\kappa\| > 1$	$\|\kappa\| = 1$	$\|\kappa\| = 1$	$\|\kappa\| = 1$	$\|\kappa\| < 1$	$\|\kappa\| = 1$	$\|\kappa\| < 1$
$\check{C} < 0$	$C < 0$	$\check{C} = 0$	$C = 0$	$C = 0$	$C = 0$	$\check{C} = 0$	$C > 0$	$\check{C} > 0$
			$\delta = \nu_\ell = \nu_r + 1$	$\delta \leq \min\{\nu_\ell, \nu_r\}$	$\delta = \nu_r = \nu_\ell + 1$			
strictly leftgoing (1)–(2)				stationary (5)			strictly rightgoing (8)–(9)	
leftgoing (1)–(5)				rightgoing (5)–(9)				

TABLE 2.2

Dispersion curve $\omega(\kappa)$	multiplicity of κ_0	$C(\kappa_0, z_0)$	Leftgoing modes	Rightgoing modes
	1	> 0	–	κ^j
	1	< 0	κ^j	–
	2	0	κ^j	κ^j
	2	0	κ^j	κ^j
	3	0	κ^j	$\kappa^j, j\kappa^j$
	3	0	$\kappa^j, j\kappa^j$	κ^j
	4	0	$\kappa^j, j\kappa^j$	$\kappa^j, j\kappa^j$
	4	0	$\kappa^j, j\kappa^j$	$\kappa^j, j\kappa^j$

2.4 Application: three-point linear multistep formulas

In this section we study the class of separable difference models of (1.1.1) with spatial discretization

$$\frac{\partial}{\partial x} \approx \frac{1}{2h}(K - K^{-1}). \tag{2.4.1}$$

These formulas have been considered previously by Beam, Warming, and Yee [Be79, Be81]. In examining them we will apply virtually all of the ideas that have been introduced so far, and in later sections they will serve repeatedly as examples. (See especially §3.2 and §6.4.)

We define these schemes by means of shift operators:

Defn. **A three-point linear multistep formula** for (1.1.1) is a separable scalar difference formula

$$\rho(Z)v_j^n = \frac{a\lambda}{2}\sigma(Z)(K - K^{-1})v_j^n, \tag{2.4.2}$$

where ρ and σ are polynomials in Z and Z^{-1}. //

The notation and terminology come from the theory of difference methods for ordinary differential equations: if (1.1.1) is discretized in space by means of (2.4.1), one obtains the system of equations

$$\frac{du_j}{dt} = \frac{a}{2h}(u_{j+1} - u_{j-1}) \qquad j \in \mathbb{Z},$$

and (2.4.2) is the fully discrete formula obtained if one solves this by a *linear multistep method* with characteristic polynomials ρ and σ [Be81].

Three of the schemes we have considered in previous sections are three-point linear multistep formulas:

$$\begin{aligned} LF: \quad & \rho(Z) = \tfrac{1}{2}(Z - Z^{-1}), & \sigma(Z) &= 1 \\ CN: \quad & \rho(Z) = Z - 1, & \sigma(Z) &= \frac{Z+1}{2} \\ BE: \quad & \rho(Z) = Z - 1, & \sigma(Z) &= Z \end{aligned}$$

The others we have looked at—LW, LF4, LFd—do not fall into this class. For further examples see [Be79].

Let us now examine the properties of a three-point linear multistep scheme that we assume to be Cauchy stable and consistent with (1.1.1).

Dispersion relation (§1.1). From (2.4.2) we obtain immediately the dispersion relation

$$\frac{\rho(z)}{\sigma(z)} = \frac{a\lambda}{2}\left(\kappa - \frac{1}{\kappa}\right), \tag{2.4.3}$$

or by (2.1.1),

$$\frac{\rho(e^{i\omega k})}{\sigma(e^{i\omega k})} = -ia\lambda \sin \xi h. \tag{2.4.4}$$

Orders of dispersion, dissipation, accuracy (§1.1). The spatial discretization (2.4.1) has order of dispersion $\alpha = 3$, order of dissipation $\beta = \infty$, and order of accuracy $\min\{\alpha, \beta\} - 1 = 2$. Except in degenerate cases (e.g. *LF with $a\lambda = 1$*), Q cannot do better than this, so it will have $\alpha = 3$ (consistency rules out $\alpha = 1$), $2 \le \beta \le \infty$, and order of accuracy 1 or 2 depending on whether β is 2 or ≥ 4. The consistency condition $\alpha, \beta \ge 2$ can also be written

$$\frac{\rho(z)}{\sigma(z)} = (z - 1) + O((z-1)^2) \quad \text{as } z \to 1. \tag{2.4.5}$$

Group velocity (§1.2). Differentiation of (2.4.3) gives

$$\left[\frac{\rho'\sigma - \rho\sigma'}{\sigma^2}\right] dz = \frac{a\lambda}{2}(1 + \kappa^{-2})\, d\kappa,$$

which by (2.3.4) gives the group speed

$$C = -\frac{a}{2}\left(\kappa + \frac{1}{\kappa}\right)\left[\frac{\sigma^2/z}{\rho'\sigma - \rho\sigma'}\right] \tag{2.4.6}$$

for any wave with $|\kappa| = |z| = 1$. From this and (2.1.1), or from (2.4.4) and (1.2.3), one obtains equivalently

$$C = -aik \cos \xi h \left[\frac{\sigma^2}{\dot\rho\sigma - \rho\dot\sigma}\right], \tag{2.4.7}$$

where $\dot\rho$ denotes $d\rho(e^{i\omega k})/d\omega$, *and similarly for* $\dot\sigma$.

Reversing properties (§1.5). If κ, z satisfy (2.4.3) with $\kappa = 1$, then the same holds with $\kappa = -1$. Moreover by (2.4.6), the latter solution has the negative of *the group* velocity of the former. Therefore Q is *x-reversing*. One cannot determine whether Q is *t-reversing* without further information on ρ and σ. (For example, LF is *t*-reversing, but CN and BE are not.)

Separability (§2.1). That Q is separable follows from the definition (2.4.2). Eqs. (2.4.6) *and* (2.4.7) *confirm* the consequence (2.1.14), that C factors into the product of a spatial and a temporal term.

Cauchy stability (§2.2). By assumption Q is Cauchy stable, which means that ρ and σ must be such that $|z| \le 1$ whenever $|\kappa| = 1$, with simple roots z for any κ, z with $|\kappa| = |z| = 1$.

Dissipativity (§2.2). By consistency (*cf.* (2.4.5)) $\kappa = z = 1$ is a solution to (2.4.3), and from (2.4.3) it follows that $\kappa = -1$, $z = 1$ is also a solution. Therefore Q cannot be *x*-dissipative, or totally dissipative. (This also follows from Thm. 2.2.4 and the fact that Q is *x*-reversing.) It can however be *t*-dissipative, depending on ρ and σ, and will *necessarily* be so if it is a two-level scheme such as CN or BE (Thm. 2.2.3).

Leftgoing and rightgoing solutions (§2.3). From (2.4.3) follows the quadratic equation for κ,

$$\kappa^2 - \frac{2\rho(z)}{a\lambda\sigma(z)}\kappa - 1 = 0,$$

and from this it is evident that for all $z \in \mathbb{C}$ there are two roots, say κ_ℓ and κ_r, satisfying

$$\kappa_\ell \kappa_r = -1. \tag{2.4.8}$$

For $|z| > 1$ these must have modulus different from 1, so we can write

$$|\kappa_r| < 1 < |\kappa_\ell| \qquad \text{for } |z| > 1, \tag{2.4.9}$$

and hence by continuity,

$$|\kappa_r| \le 1 \le |\kappa_\ell| \qquad \text{for } |z| \ge 1, \tag{2.4.10}$$

The subscripts ℓ and r refer to "leftgoing" and "rightgoing", respectively; in fact (2.4.9) *implies* that the waves $\kappa_\ell^j z^n$ and $\kappa_r^j z^n$ are strictly left- and rightgoing, respectively, for $|z| > 1$. For $|z| = 1$ the strictness will be lost if $|\kappa_\ell| > 1$ and $|\kappa_r| < 1$, but it will be *preserved* if $|\kappa_\ell| = |\kappa_r| = 1$, unless $C = 0$, which by Thm. 2.3.1 and (2.4.8) will happen if and *only* if $\kappa_r = \kappa_\ell = \pm i$. In any case there is exactly one leftgoing value $\kappa_\ell(z)$ and one rightgoing value $\kappa_r(z)$, continuously defined for $|z| \ge 1$.

• • •

In *many* cases where one picks, say, LF to illustrate a point about difference models, it is really its spatial discretization that the illustration depends on, and any three-point linear multistep *formula* will show the same thing. With this in mind we have described this class partly to avoid having to present future examples in too limited a context.

For some applications we will be interested in two subclasses of the set of three-point linear multistep formulas. The following definitions are derived from the theory of linear multistep methods for ordinary differential equations:

Defn [Be81]. Let Q be a three-point linear multistep formula consistent with (1.1.1). We say that Q is ***A*-stable** if it is Cauchy stable and has the property

(i) $\mathrm{Re}\, a(\kappa - 1/\kappa) \leq 0 \Rightarrow |z| \leq 1$, with z simple if $|z| = 1$. (2.4.11)

Q is **strongly *A*-stable** if (i) holds and furthermore

(ii) $\mathrm{Re}\, a(\kappa - 1/\kappa) \leq 0,\ \kappa \neq \pm 1 \Rightarrow |z| < 1$, (2.4.12)

(iii) $\rho(z) = 0,\ |z| = 1 \Rightarrow z = 1$ (cf. 2.4.5) (2.4.13)

The motivation for these definitions, which is discussed in most books on the numerical solution of ordinary differential equations, is that they provide conditions for Q to be stable for arbitrarily large mesh ratios λ. Beam et al. have pointed out that this is a desirable property if one wishes to apply a time-dependent difference model to find the steady-state solution of a physical problem, without being concerned about the accuracy for the transient computation (see §6.4).

A-stable schemes have some simple properties that will turn out to be important to their stability analysis:

Theorem 2.4.1. *Let Q be a three point linear multistep formula consistent with $u_t = au_x$ with $a > 0$.*

(i) If Q is A-stable, then

$$Re\,\kappa_r \leq 0 \leq Re\,\kappa_\ell \tag{2.4.14}$$

for all z with $|z| \geq 1$.

(ii) If Q is strongly A-stable, then

$$Re\,\kappa_r < 0 < Re\,\kappa_\ell \tag{2.4.15}$$

and

$$|\kappa_r| < 1 < |\kappa_\ell| \tag{2.4.16}$$

for all z with $|z| \geq 1$, except when $\kappa_\ell = -\kappa_r = \pm 1$.

If $a < 0$, the same results hold with the inequalities in (2.4.14) and (2.4.15) reversed.

Proof. Assume $a > 0$; the proofs for $a < 0$ are similar.

If Q is A-stable, then the contrapositive to (2.4.11) asserts that for $|z| > 1$, one has $\mathrm{Re}(\kappa - 1/\kappa) > 0$. Taking $\kappa = \kappa_\ell$ and using (2.4.9), one obtains $\mathrm{Re}\,\kappa_\ell > 0$. With (2.4.8) this implies $\mathrm{Re}\,\kappa_r < 0 < \mathrm{Re}\,\kappa_\ell$ for $|z| > 1$, and (2.4.14) follows by continuity.

If Q is strongly A-stable, then the contrapositive to (2.4.12) implies further that for $|z| \geq 1$, either $\kappa = \pm 1$ or $\mathrm{Re}(\kappa - 1/\kappa) > 0$. Together with (2.4.8), (2.4.10), and (2.4.14), the latter formula implies (2.4.15) and (2.4.16), as required. ∎

So far we have not used condition (2.4.13), but it has a simple consequence:

Theorem 2.4.2. *Let Q be a three-point linear multistep formula for (1.1.1). If Q is strongly A-stable, then it is t-dissipative.*

Proof. Suppose $\kappa = 1$ and $|z| = 1$. The first of these conditions implies $\rho(z) = 0$ by (2.4.3), and by (2.4.13), the second then implies $z = 1$. This establishes t-dissipativity. ∎

2.5 Extension from scalars to diagonalisable systems

In practice one is generally concerned not with one scalar equation, but with a **hyperbolic system of equations**. Such a system takes the form

$$u_t = Au_x, \tag{2.5.1}$$

where $u(x,t)$ is an N-vector and A is a square matrix of dimension N. For simplicity we assume as before that A is constant, and we continue to omit any undifferentiated terms.

Following (2.1.3), we can write a general constant coefficient model Q of (2.5.1) in the form

$$Q_{-1}v_j^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma v_j^{n-\sigma}. \tag{2.5.2}$$

Now each v_j^n is an N-vector, and each Q_σ is a constant spatial difference operator with square matrix coefficients of dimension N. If these coefficients are denoted by $A_{j\sigma}$, then the analog of (2.1.4) becomes

$$Q_\sigma = \sum_{j=-\ell}^{r} A_{j\sigma} K^j. \tag{2.5.3}$$

matrix coefficients,

$$P(K,Z)v = \left[\sum_{j=-\ell}^{r}\sum_{\sigma=-1}^{s} A_{j\sigma}K^{j+\ell}Z^{s-\sigma}\right]v = 0. \qquad (2.5.4)$$

If the system (2.5.1) is hyperbolic, then A can be diagonalized and it has real eigenvalues. In principle, the matrices $\{A_{j\sigma}\}$ might not have this property, or they might each be diagonalizable without the existence of a single matrix to diagonalize all of them simultaneously. But this rarely happens in practice, and indeed usually each $A_{j\sigma}$ is a polynomial in A, so they are all diagonalized by the same matrix as A. Therefore we will make the asssmption (= Ass. 5.4 of [Gu72]):

Assumption 2.1. The matrices $\{A_{j\sigma}\}$ are simultaneously diagonalizable. That is, there exists a constant nonsingular $N \times N$ matrix T such that

$$\tilde{A}_{j\sigma} = TA_{j\sigma}T^{-1} = \mathrm{diag}(a_{j\sigma}^{(1)},\dots,a_{j\sigma}^{(N)}), \qquad (2.5.5)$$

with $a_{j\sigma}^{(\alpha)} \in \mathbb{R}$ for all α, j, σ. //

With this assumption, the study of wave propagation under difference models of (2.5.1) reduces directly to the results already established for scalar problems. From (2.5.4) and (2.5.5), one obtains

$$\tilde{P}(K,Z)\tilde{v} = \left[\sum_{j=-\ell}^{r}\sum_{\sigma=-1}^{s} \tilde{A}_{j\sigma}K^{j+\ell}Z^{s-\sigma}\right]\tilde{v} = 0, \qquad (2.5.6)$$

where $\tilde{v}$ denotes Tv and $\tilde{P}$ denotes TPT^{-1}. Now $\tilde{P}$ is a bivariate polynomial with diagonal matrix coefficients. This system is equivalent to the N scalar systems

$$\tilde{P}^{(\alpha)}(K,Z)\tilde{v}^{(\alpha)} = \left[\sum_{j=-\ell}^{r}\sum_{\sigma=-1}^{s} a_{j\sigma}^{(\alpha)}K^{j+\ell}Z^{s-\sigma}\right]\tilde{v}^{(\alpha)}, \qquad 1 \le \alpha \le N. \qquad (2.5.7)$$

Each equation (2.5.7) has the same form as (2.1.5). Corresponding to the polynomials P_z and P_κ of §2.1, we can also define matrix polynomials P_z, P_κ, $\tilde{P}_z$, and $\tilde{P}_\kappa$ in the obvious way, and $\tilde{P}_z$ and $\tilde{P}_\kappa$ are diagonal with scalar components $\tilde{P}_z^{(\alpha)}$ and $\tilde{P}_\kappa^{(\alpha)}$.

Following (2.1.18), we now ask: given $z \in \mathbb{C}$, what solutions of the form

$$v_j^n = z^n\Phi_j, \qquad (2.5.8)$$

where $\{\Phi_j\}$ is a sequence of N-vectors, does Q support? Such solutions will be precisely those sequences satisfying, in extension of (2.1.10), the matrix resolvent equation

$$P_z(K)\Phi_j = P(K,z)\Phi_j = 0. \qquad (2.5.9)$$

The following theorem is an extension of Thm. 2.1.2 (cf. [Gu72], eq. (5.5)):

Theorem 2.5.1. *Let Q satisfy Assumption 2.1, and let z satisfy $|z| \ge 1$. For $1 \le \alpha \le N$ let $\{\kappa_i^{(\alpha)}\}_{1\le i\le \mu^{(\alpha)}}$ denote the distinct nonzero roots of $\tilde{P}_z^{(\alpha)}$, with $\kappa_i^{(\alpha)}$ of multiplicity $\nu^{(\alpha)}$. Then the sequences*

$$\Phi_j = [\kappa_i^{(\alpha)}]^j j^\delta \psi^{(\alpha)} \qquad \begin{matrix} 1 \le \alpha \le N \\ 1 \le i \le \mu^{(\alpha)} \\ 0 \le \delta \le \nu_i^{(\alpha)} - 1 \end{matrix} \qquad (2.5.10)$$

are linearly independent solutions of (2.5.9), and they span the linear space of all such solutions. Here $\psi^{(\alpha)}$ denotes $T^{-1}(0,\dots,0,1,0,\dots,0)^T$, where the 1 is in position α.

Proof. Diagonalization of (2.5.9) by T gives $\tilde{P}_z(K)\tilde{\Phi} = 0$ with $\tilde{\Phi} = T\Phi$. The solutions to this equation are given componentwise by Thm. 2.1.2, and have the form $\tilde{\Phi}_j = [\kappa_i^{(\alpha)}]^j j^\delta(0,\dots,0,1,0,\dots,0)^T$. Multiplying by T^{-1} completes the proof. ∎

This theorem completely describes the solutions with regular behavior in t that are admitted by Q. Each one is nothing more than a scalar signal transformed to the basis determined by T. Therefore all of the theory derived earlier applies directly. For $\delta = 0$ and $|z| = |\kappa_i^{(\alpha)}| = 1$, $v_j^n = z^n\Phi_j$ represents a wave that propagates uniformly at the group velocity (cf. (2.3.4))

$$C_i^{(\alpha)} = \frac{d\omega}{d\xi_i^{(\alpha)}} = -\frac{1}{\lambda}\left(\frac{dz}{d\kappa_i^{(\alpha)}}\right)\left(\frac{\kappa_i^{(\alpha)}}{z}\right). \qquad (2.5.11)$$

We say that the signal v_j^n is **leftgoing, rightgoing, strictly leftgoing, strictly rightgoing,** or **stationary** precisely when the corresponding terms hold for the scalar signal $z^n[\kappa_i^{(\alpha)}]^j j^\delta$.

The definitions of the **von Neumann condition** and **Cauchy stability** given in §2.2 apply as written to the vector model Q. (The symbol $|\phi_j|$ in the definition of the latter must be interpreted as the two-norm of N-vectors rather than an absolute value.) It follows from these definitions that Q satisfies the von Neumann condition, or is Cauchy stable, precisely when the same is true for all of the scalar problems in the diagonalization (2.5.7).

Let n_ℓ and n_r denote the total number of linearly independent leftgoing and rightgoing signals, respectively, admitted by Q for some z with $|z| \ge 1$. Then by

(2.5.10), the general solution of the form (2.5.8) can be written

$$v_j^n = \sum_{i=1}^{n_r} a_i \kappa_i^j j^{\ell_i} \psi_i + \sum_{i=n_r+1}^{n_r+n_\ell} a_i \kappa_i^j j^{\ell_i} \psi_i. \tag{2.5.12}$$

It is obvious that Assumption 2.1 has rendered the developments of this section fairly trivial, and one may wonder why it is worth mentioning systems of equations at all if they are only to be reduced immediately to scalars. The answer is that as we turn to calculations of reflection and transmission coefficients, and then to stability for initial boundary value problems, boundary terms will appear that couple the scalar components together and cannot be diagonalised away. The meaning of this for practical applications is that although a hyperbolic system of equations can be reduced to *characteristic variables* in the interior, it may be desired to give boundary conditions in terms of *primitive variables*. For more on this distinction see [Co80] and [Gu82].

• • •

Let us finish the section with a simple example of a difference model for a system of hyperbolic equations (cf. §5.1 of [Co80] and §4 of [Gu75]).

Example 2.1. Let the hyperbolic system

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{bmatrix} a & 1 \\ 1 & a \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x \tag{2.5.13}$$

be modeled by the vector leap frog scheme

$$\begin{pmatrix} u_j^{n+1} \\ v_j^{n+1} \end{pmatrix} = \begin{pmatrix} u_j^{n-1} \\ v_j^{n-1} \end{pmatrix} + \lambda \begin{bmatrix} a & 1 \\ 1 & a \end{bmatrix} \left(\begin{pmatrix} u_{j+1}^n \\ v_{j+1}^n \end{pmatrix} - \begin{pmatrix} u_{j-1}^n \\ v_{j-1}^n \end{pmatrix} \right), \tag{2.5.14}$$

where we have abused notation by using the letters u, v for both exact and computed variables. Eq. (2.5.13) can be diagonalized by the matrix

$$T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

which converts it to

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix}_t = \begin{bmatrix} a-1 & 0 \\ 0 & a+1 \end{bmatrix} \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix}_x,$$

with

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = T \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u - v \\ u + v \end{pmatrix}.$$

Thus $u - v$ and $u + v$ are the characteristic variables for (2.5.13). The same matrix T diagonalises (2.5.14), and therefore the vector leap frog model decouples into LF for each of the two scalar problems $\tilde{u}_t = (a-1)\tilde{u}_x$ and $\tilde{v}_t = (a+1)\tilde{v}_x$. It follows that for any z with $|z| \geq 1$, (2.5.14) admits four fundamental solutions (2.5.10), namely

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} [\kappa^{(1)}]^j z^n, \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix} \left[\frac{-1}{\kappa^{(1)}} \right]^j z^n,$$

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} [\kappa^{(2)}]^j z^n, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left[\frac{-1}{\kappa^{(2)}} \right]^j z^n.$$

If $|z| = |\kappa^{(\alpha)}| = 1$, the first two have equal and opposite group velocities in the range between $\pm|a-1|$, and the latter two have equal and opposite group velocities in the range $\pm|a+1|$.

3. BOUNDARIES AND INTERFACES

3.1 Reflection and transmission coefficients

Most practical finite difference models are complicated by the presence of boundaries or interfaces, at which the properties of the model change abruptly with respect to x. A boundary may be imposed physically by the problem being modeled, or it may be a numerical artifact required to keep the grid finite [En77]. Likewise an interface may represent a discontinuity in the physical medium [Br79,Ma81,Su74], or a numerical discontinuity such as a change of mesh size (**mesh refinement**) [Ci71,Br73,Vi81b] or of difference formula (**hybridization**) [Ci72,Ol76]. Also, if the solution to a partial differential equation contains shocks or other discontinuities, it may be useful to think of these as moving interfaces [Ap68,Ch78,Ch79]. Whether a boundary or interface is physical or purely numerical does not affect the procedure for analyzing its numerical behavior, which we will describe in this chapter. Of course it does affect the results of this analysis and their interpretation. *For example*, a physical boundary or interface may be expected to reflect some energy backwards when a wave strikes it, even for $\omega, \xi \approx 0$, whereas any energy reflected by a purely numerical interface is spurious, and must approach 0 for $\omega = \xi = 0$.

Our approach to the analysis of reflection and transmission problems is based on the examination of ***steady-state solutions*** with regular behavior z^n in t. On the face of it this is Fourier analysis with respect to t, but the subtlety of the problem comes from the inevitable need to make a connection between the Fourier spectrum in t and that in x. Fundamental to this connection is the distinction between ***leftgoing*** and ***rightgoing*** solutions presented in Chapter 2. *In §§3.1–3.4 we study scalar monochromatic signals*, and in §3.5 we superpose these to consider reflection of a general wave packet. In §3.6 the formulation is generalized from scalars to diagonalizable systems, and we introduce a general notation for reflection problems.

Here is the main idea. Suppose that the wave front of a monochromatic wave

$e^{i(\omega t-\xi_0 x)}$ with $\omega, \xi_0 \in \mathbb{R}$, or more generally of any signal $z^n\kappa_0^j$ (2.1.2) with $\kappa_0, z \in \mathbb{C}$ and $|z| \geq 1$, hits a boundary or interface from one side. The interaction will be complicated at first. As t increases, however, a **steady-state solution** will normally be approached in which the incident signal is balanced by a collection of monochromatic reflected and possibly transmitted signals $z^n\kappa_\iota^j j^{\delta_\iota}$. All of these signals will have *the same time variation factor* z, but their space factors κ_ι will vary. For the case of an interface at $j = j_0$, with the incident wave coming from the left, the steady-state solution will take the form (cf. Thm. 2.1.2)

$$v_j^n = \begin{cases} z^n\kappa_0^j + \sum_{\iota\in I_\ell} a_\iota z^n \kappa_\iota^j j^{\delta_\iota} & j \leq j_0, \\ \sum_{\iota\in I_r} a_\iota z^n \kappa_\iota^j j^{\delta_\iota} & j \geq j_0. \end{cases} \tag{3.1.1}$$

Here I_ℓ and I_r are "left" and "right" index sets, *respectively*. In this notation a κ value of multiplicity ν appears ν times in the index set, with corresponding δ values $0, \dots, \nu-1$. The modifications of (3.1.1) for incidence from the right, or for a boundary instead of an interface, are obvious. Depending on labeling of points, the precise form of the solution might also change in unimportant ways for $j \approx j_0$.

Two principles determine what κ's may appear in (3.1.1):

The set $\{(\kappa_\iota, \delta_\iota)\}$ indexed by I_ℓ (resp. I_r) consists of precisely those distinct pairs $(\kappa_\iota, \delta_\iota)$ for which:

(1) (κ_ι, z) satisfies the dispersion relation for the difference formula applied in $j < j_0$ (resp. $j > j_0$), with κ_ι of multiplicity $\nu \geq \delta_\iota$ (Thm. 2.1.2); and

(2) The signal (2.3.6) with parameters $\kappa_\iota, z, \delta_\iota$ is leftgoing (resp. rightgoing) (see Table 2.1).

The interesting restriction is (2), for it shows that the numerical behavior of boundaries *and interfaces* depends fundamentally on group velocity. The principle is simple: a wave impinging on the interface can stimulate only energy that propagates outward from the interface, not energy coming in *from infinity*. *In physics* this is called the **Sommerfeld radiation condition**. We will not attempt to justify the condition mathematically in the sense of showing that transient signals approach (3.1.1) as $t \to \infty$. By construction, however, (3.1.1) is itself guaranteed to be a solution of the difference model.

We emphasize that the signals present in I_ℓ and I_r are determined by numerical wave behavior entirely, so they may be any mix of physically realistic waves, parasites, or signals in between. For $|z| = 1$ *some* may have $|\kappa| = 1$, and others $|\kappa| \ll 1$ or

$|\kappa| \gg 1$. Only the amplitudes $\{a_i(z)\}$ of the stimulated signals are affected by the algebraic details at the interface, and determining them will be a matter of linear algebra. These amplitudes, one for each outgoing signal, are the **reflection and transmission coefficients** for the given problem.

For setting up interface conditions we need to rule out possible degeneracies in the difference model. We will assume that the difference formulas Q appearing on either side of the interface satisfy the following condition (cf. Ass. 5.5 of [Gu72]).

Assumption 3.1. Q is Cauchy stable, and for all z with $|z| \geq 1$, the polynomial $P_z(\kappa)$ of §2.1 has nonzero 0th and $(\ell+r)$th coefficients. Moreover, of the $\ell+r$ solutions (2.3.6) admitted by Q, exactly r are leftgoing and exactly ℓ are rightgoing. //

We will let the symbol $\tilde{Q}$ denote the complete difference model, consisting of one or more "interior" difference formulas Q, Q_-, Q_+, etc. together with additional conditions imposed at the boundary or interface.

3.2 Examples

The best way to show how reflection and transmission coefficients are calculated is through examples. We will now give a number of these, deferring a more formal treatment to §3.6, and in the process explore various problems of interest in their own right. Most of the results derived here will be applied in later sections.

Example 3.1: LF with abrupt coefficient change

Consider a first-order equation with discontinuous coefficients,

$$u_t = \begin{cases} a_- u_x & (x < 0) \\ a_+ u_x & (x > 0) \end{cases} \qquad a_-, a_+ \neq 0. \tag{3.2.1}$$

If $a_-, a_+ < 0$, the solutions to this equation consist of rightgoing waves, which pass through $x = 0$ with no alteration but a change in wave number. In particular, no energy is reflected backwards. However, let (3.2.1) be modeled by LF (1.1.6) on the grid $x_j = jh$ for $j = \ldots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots$, with $a_j \equiv a_-$ for $j \leq -\frac{1}{2}$ and $a_j \equiv a_+$ for $j \geq \frac{1}{2}$. Now, when a smooth wave passes rightward through the interface, a leftgoing reflected parasite will be generated. If $a_-, a_+ > 0$, on the other hand, then a sawtoothed wave can travel rightwards through the interface, and it will generate a reflected signal of low wave number.

Let us determine the steady-state configuration that results when a strictly rightgoing signal $\kappa_i^j z^n$ (2.1.2) with $|z| \geq 1$ hits the interface from the left. Whatever the signs of a_- and a_+, there are three signals to consider: one incident, one transmitted, and one reflected. Their functional forms are indicated in Fig. 3.1:

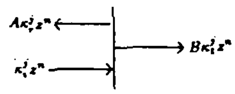


FIG. 3.1

The j's in these expressions are half-integers. We will ignore the question of the choice of square roots; it does not affect the final result.

Given z, the quantities $\kappa_i, \kappa_t, \kappa_r$ are determined by the dispersion relation (2.1.8) on the left and right:

$$z - \frac{1}{z} = \lambda a_-(\kappa_i - \kappa_i^{-1}) = \lambda a_-(\kappa_r - \kappa_r^{-1}) = \lambda a_+(\kappa_t - \kappa_t^{-1}). \tag{3.2.2}$$

Our purpose is to find the reflection and transmission coefficients A and B. The equations needed to determine them are the "interface formulas" at $j = \pm\frac{1}{2}$, which assert that the steady-state solution satisfies the difference formulas at those points:

$$v_{-1/2}^{n+1} - v_{-1/2}^{n-1} = \lambda a_-(v_{1/2}^n - v_{-3/2}^n),$$
$$v_{1/2}^{n+1} - v_{1/2}^{n-1} = \lambda a_+(v_{3/2}^n - v_{-1/2}^n).$$

Inserting the wave forms of Fig. 3.1 in these equations gives

$$\begin{aligned} (z - \tfrac{1}{z})(\kappa_i^{-1/2} + A\kappa_r^{-1/2}) &= \lambda a_-(B\kappa_t^{1/2} - \kappa_i^{-3/2} - A\kappa_r^{-3/2}), \\ (z - \tfrac{1}{z})B\kappa_t^{1/2} &= \lambda a_+(B\kappa_t^{3/2} - \kappa_i^{-1/2} - A\kappa_r^{-1/2}). \end{aligned} \tag{3.2.3}$$

We could solve these equations for A and B and get formulas involving $z, \kappa_i, \kappa_r, \kappa_t, a_-$, and a_+. In general, this is the best that can be done. However, for simple problems one may conveniently eliminate z. In the present case, applying (3.2.2) to (3.2.3) eliminates not only z but $a_\pm$ as well, leaving

$$\kappa_i^{-1/2}(\kappa_i - \kappa_i^{-1}) + A\kappa_r^{-1/2}(\kappa_r - \kappa_r^{-1}) = B\kappa_t^{1/2} - \kappa_i^{-3/2} - A\kappa_r^{-3/2},$$
$$(\kappa_t - \kappa_t^{-1})B\kappa_t^{1/2} = B\kappa_t^{3/2} - \kappa_i^{-1/2} - A\kappa_r^{-1/2},$$

hence because of cancellations,

$$B\kappa_t^{1/2} = \kappa_i^{1/2} + A\kappa_r^{1/2},$$
$$B\kappa_t^{-1/2} = \kappa_i^{-1/2} + A\kappa_r^{-1/2}. \tag{3.2.4}$$

The solution to this pair of equations is

$$A = -\frac{\kappa_t - \kappa_i}{\kappa_t - \kappa_r}\sqrt{\frac{\kappa_r}{\kappa_i}}, \qquad B = \frac{\kappa_r - \kappa_i}{\kappa_r - \kappa_t}\sqrt{\frac{\kappa_t}{\kappa_i}}. \tag{3.2.5}$$

We have now solved the reflection and transmission problem: given z, first compute κ_i, κ_r, κ_t from (3.2.2), then derive A and B from (3.2.5).

Eqs. (3.2.5) have a pleasing symmetry that becomes particularly useful in the case of strictly wavelike solutions, i.e. $|z| = |\kappa_i| = |\kappa_t| = |\kappa_r| = 1$. Let us write

$$\kappa_i = e^{-i\xi_i h} = e^{-2i\theta_i}, \quad \kappa_t = e^{-i\xi_t h} = e^{-2i\theta_t}, \quad \kappa_r = e^{-i\xi_r h} = e^{-2i\theta_r}.$$

Then (3.2.5) is readily seen to take the form

$$A = -\frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t - \theta_r)}, \qquad B = \frac{\sin(\theta_r - \theta_i)}{\sin(\theta_r - \theta_t)}. \tag{3.2.6}$$

(Of course, these formulas are also valid for $\theta \notin \mathbb{R}$.)

A further simplification follows from the fact that for LF, κ_i and κ_r or θ_i and θ_r are related in a simple way. From (2.4.8) one has $\kappa_r = -1/\kappa_i$, hence

$$\theta_r = \frac{\pi}{2} - \theta_i. \tag{3.2.7}$$

With these substitutions (3.2.5) and (3.2.6) become

$$A = \frac{1}{i}\,\frac{\kappa_t - \kappa_i}{\kappa_i\kappa_t + 1}, \qquad B = \frac{\kappa_i + 1/\kappa_i}{\kappa_i\kappa_t + 1}\sqrt{\kappa_i\kappa_t}, \tag{3.2.8}$$

$$A = \frac{\sin(\theta_t - \theta_i)}{\cos(\theta_t + \theta_i)}, \qquad B = \frac{\cos 2\theta_i}{\cos(\theta_t + \theta_i)}. \tag{3.2.9}$$

These equations show that in the limit of a vanishing interface, i.e. $a_- \approx a_+$ and hence $\theta_i \approx \theta_t$, *one obtains the physically correct values $A \approx 0$, $B \approx 1$. In fact they imply $A = O(a_+ - a_-)$ as $a_+ - a_- \to 0$.*

DEMONSTRATION 3.1. Fig. 3.2 shows an experiment with $a_- = -1$, $a_+ = -.5$,

$\lambda = .5$, $h = .01$ on the interval $[-1, 1]$. At $t = 0$ the oscillation

$$u(-1, t) = \sin 30t$$

has been turned on. This generates a rightgoing wave that is well resolved on the mesh (≈ 21 points per wavelength), and Fig. 3.2a shows that by $t = .5$, it has traveled at the correct group speed $C \approx 1$ and should hit $x = 0$ at $t \approx 1$. In Fig. 3.2b, showing $t = 1.5$, it is evident that the transmitted wave must have traveled at approximately its correct speed $C = \frac{1}{2}$. We are interested in the reflected parasitic wave that appears as wiggles in the region $[-1, 0]$. Apparently it has moved at speed $C \approx -1$, which is

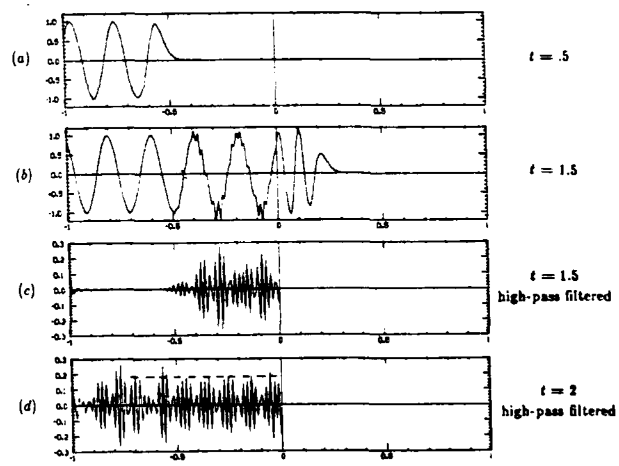


FIG. 3.2. Reflection and transmission at an LF interface. A forcing oscillation with $\omega k = .15$ has been turned on at $t = 0$ and hits a coefficient-change interface at $t \approx 1$. The model is LF for $u_t = -u_x$ on the left, $u_t = -.5u_x$ on the right, with $h = 1/100$, $\lambda = .5$.

what we expect for LF. From (3.2.9) we can predict its amplitude. We have

$$\theta_i = \xi_i h/2 \approx \omega h/2 = .15,$$

$$\theta_t = \xi_t h/2 \approx 2\xi_i h/2 \approx .30.$$

Eq. (3.2.9) therefore gives

$$A \approx \frac{\sin .15}{\cos .45} \approx .166.$$

The exact value for (3.2.9) turns out to be $A = .1884\ldots$. It is hard to tell from Fig. 3.2b how well this agrees with the amplitude of the wiggles in the experiment. Therefore Fig. 3.2c isolates these wiggles by showing the result of passing the function in $[-1, 0]$ of Fig. 3.2b through a high-pass filter (discrete Fourier transform; zeroing of lower half of spectrum; inverse transform). Fig. 3.2d gives a similar filtered plot for $t = 2$, after the initial transients in the reflected wave have died down. The agreement with the prediction $|A| = .1884$, represented by the dashed line in Fig. 3.2d, is obviously excellent.

Example 3.2: Abrupt change between arbitrary 3-point schemes

Consider eqs. (3.2.4) of the last example. Although our derivation made use of the dispersion relation for LF, it is obvious that what these equations really assert is this: at $j = -1/2$ and at $j = 1/2$, the lefthand representation $v_j^n = (\kappa_i^j + A\kappa_r^j)z^n$ and the righthand representation $v_j^n = B\kappa_t^j z^n$ are *both valid.* (A priori, we knew only that the former was valid at $j = -1/2$ and the latter at $j = 1/2$ (Fig. 3.1).) This suggests that the calculations of Example 3.1 have a wider applicability. This is in fact the case.

Let Q_- and Q_+ then be arbitrary three-point difference formulas as described in §2.1, to be applied for $j \le -1/2$ and $j \ge 1/2$, respectively. By this we mean that the stencils satisfy $\ell_- = r_- = 1$, $\ell_+ = r_+ = 1$. (In fact all we need is $\ell_+ = r_- = 1$.) Assume further that Assmption 3.1 holds. The following argument shows that eqs. (3.2.4) must hold. We know that the representation $v_j^n = B\kappa_t^j z^n$ is valid for $j \ge 1/2$. By the definition of κ_t, it follows that if $v_{-1/2}^n = B\kappa_t^{-1/2} z^n$ also, then Q_+ will be satisfied at $j = 1/2$. But Q_+ is satisfied there, and Ass. 3.1 implies that if the values v_j^n for $j \ge 1/2$ are fixed, then this can only happen for a unique value of $v_{-1/2}^n$. Therefore $v_{-1/2}^n = B\kappa_t^{-1/2} z^n$. The result at $j = 1/2$ is similar.

Thus most of the calculations of Example 3.1 apply not just to (3.2.1) modeled by LF, but to any interface at which one three-point difference formula changes abruptly

to another. The interface may involve just a change of coefficient, as before, or it may include a change of difference formula also, for example from LF to LW. For any such problem, one is led by (3.2.4) to (3.2.5) and (3.2.6). After this, eqs. (3.2.8) and (3.2.9) are not universally valid, but since all they depend upon is (2.4.8), they will hold whenever Q_- is a three-point linear multistep formula.

As an example, suppose (3.2.1) is replaced by the *second*-order wave equation

$$u_{tt} = \begin{cases} a_-^2 u_{xx} & (x < 0) \\ a_+^2 u_{xx} & (x > 0) \end{cases} \qquad a_-, a_+ \neq 0 \tag{3.2.10}$$

modeled by the leap frog scheme LF²,

$$v_j^{n+1} - 2v_j^n + v_j^{n-1} = (\lambda a_\pm)^2 (v_{j+1}^n - 2v_j^n + v_{j-1}^n). \tag{3.2.11}$$

This formula has the dispersion relation

$$z - 2 + z^{-1} = (\lambda a_\pm)^2 (\kappa - 2 + \kappa^{-1}),$$

from which one may see that instead of (2.4.8) and (3.2.7), κ_i and κ_r now satisfy

$$\kappa_i \kappa_r = 1, \qquad \theta_r = -\theta_i. \tag{3.2.12}$$

Now both the incident wave and the reflected wave can be physical (smooth) at the same time, for (3.2.10) permits wave motion in both directions. The reflection and transmission coefficients for (3.2.10) can be obtained by enforcing C^1 continuity at $x = 0$, and are independent of ω and ξ (see e.g. [Cl76], §8.1):

$$A = \frac{1/a_- - 1/a_+}{1/a_- + 1/a_+}, \qquad B = \frac{2/a_-}{1/a_- + 1/a_+}. \tag{3.2.13}$$

These formulas are written in a standard form in terms of the *admittances* $1/a_\pm$; one could also use the *impedances* $a_\pm$ directly. For the LF² model, the corresponding results are by (3.2.5), (3.2.6), and (3.2.12),

$$A = \frac{\kappa_i - \kappa_t}{\kappa_i \kappa_t - 1}, \qquad B = \frac{\kappa_i - 1/\kappa_i}{\kappa_i \kappa_t - 1}\sqrt{\kappa_i \kappa_t}, \tag{3.2.14}$$

or

$$A = \frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}, \qquad B = \frac{\sin 2\theta_i}{\sin(\theta_i + \theta_t)}. \tag{3.2.15}$$

This last pair of formulas is a trigonometric analog of the admittance formulas (3.2.13), and approaches them for small ξ and ω, but it is not the same.

Our calculations apply to dissipative schemes also. Let (3.2.1) be modeled, say, by LW (1.1.11). For $a_-, a_+ < 0$, a physical signal will then have $\kappa_i, \kappa_t \approx 1$ and (it follows from (2.1.9)) $|\kappa_r| \approx |\frac{\lambda a - 1}{\lambda a + 1}| > 1$. Therefore the reflected wave is evanescent, and will have negligible amplitude except near the interface. It can by no means be ignored in computing B, however, for it need not be negligible at $x = 0$—i.e. A itself need not be small. This situtation is typical for both dissipative and nondissipative models: evanescent modes are often present that have negligible size away from the interface, but their influence is still global because they affect the amplitudes of the non-evanescent modes.

Example 3.3: Abrupt change between schemes with larger stencils

The principles of Example 3.2 apply directly to difference schemes with larger stencils. Let Q_- and Q_+ have stencil sizes ℓ_-, r_- and ℓ_+, r_+, and assume that both formulas satisfy Assumption 3.1. We seek the reflected waves that result after an incident signal $\kappa_0^j z^n$ with $|z| \geq 1$ hits $j = 0$ from the left. For $j < 0$ there are r_- leftgoing signals, and if we denote their amplitudes by $-A_1, \ldots, -A_{r_-}$, these may be written

$$-A_\nu \kappa_\nu^j z^n, \qquad 1 \leq \nu \leq r_-.$$

(We ignore the possibility of defective modes.) For $j > 0$ there are ℓ_+ rightgoing signals, and we denote their amplitudes by $A_{r_-+1}, \ldots, A_{r_-+\ell_+}$:

$$A_\nu \kappa_\nu^j z^n, \qquad r_- + 1 \leq \nu \leq r_- + \ell_+.$$

Exactly as in the last example, Assumption 3.1 implies that the righthand representation of v_j^n must hold not just for $j \geq 1/2$, but for $j \geq 1/2 - \ell_+$. This follows by the same argument as before by considering in succession $j = -1/2, -3/2, \ldots, 1/2 - \ell_+$. Likewise, the lefthand representation must hold for all $j \leq -1/2 + r_-$. All together, there are $\ell_+ + r_-$ matching conditions in extension of (3.2.4), and they take the form of a van der Monde system of equations:

$$\begin{pmatrix} \kappa_1^{\frac12-\ell_+} & \kappa_2^{\frac12-\ell_+} & \cdots & \kappa_{\ell_++r_-}^{\frac12-\ell_+} \\ \kappa_1^{\frac32-\ell_+} & \kappa_2^{\frac32-\ell_+} & \cdots & \kappa_{\ell_++r_-}^{\frac32-\ell_+} \\ \vdots & \vdots & & \vdots \\ \kappa_1^{-\frac12+r_-} & \kappa_2^{-\frac12+r_-} & \cdots & \kappa_{\ell_++r_-}^{-\frac12+r_-} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_{\ell_++r_-} \end{pmatrix} = \begin{pmatrix} \kappa_0^{\frac12-\ell_+} \\ \kappa_0^{\frac32-\ell_+} \\ \vdots \\ \kappa_0^{-\frac12+r_-} \end{pmatrix}. \tag{3.2.16}$$

The determinant of this matrix is

$$\prod_{\substack{\mu,\nu=1\\ \mu>\nu}}^{\ell_++r_-} (\kappa_\mu - \kappa_\nu) \Big/ \prod_{\nu=1}^{\ell_++r_-} \kappa_\nu^{\ell_+-\frac12}.$$

According to Cramer's rule, the solution to (3.2.16) can be expressed in terms of ratios of such determinants. We find (cf. (3.2.5) and (3.2.6)):

$$A_\nu = \prod_{\substack{\mu=1\\ \mu\neq\nu}}^{\ell_++r_-} \frac{(\kappa_\mu - \kappa_0)/\kappa_0^{\ell_+-\frac12}}{(\kappa_\mu - \kappa_\nu)/\kappa_\nu^{\ell_+-\frac12}} = \prod_{\substack{\mu=1\\ \mu\neq\nu}}^{\ell_++r_-} \frac{\sin(\theta_\mu - \theta_0)}{\sin(\theta_\mu - \theta_\nu)} \left(\frac{\kappa_\nu}{\kappa_0}\right)^{\frac12(\ell_+-r_-)}. \tag{3.2.17}$$

These formulas give the complete solution of the reflection and transmission problem. In practice, if the incident signal is wavelike ($|z| = |\kappa_0| = 1$), then often some reflected and transmitted signals will be wavelike, others evanescent. However, this distinction affects the values of $\{\kappa_\nu\}$ and $\{\theta_\nu\}$, not the form of (3.2.17).

DEMONSTRATION 3.2. As a particular example, let us again consider the problem of Example 3.1, but with LF replaced by LF4 (1.1.17), whose stencil covers five grid points in x. Eq. (3.2.16) now becomes a system of dimension 4, and for typical

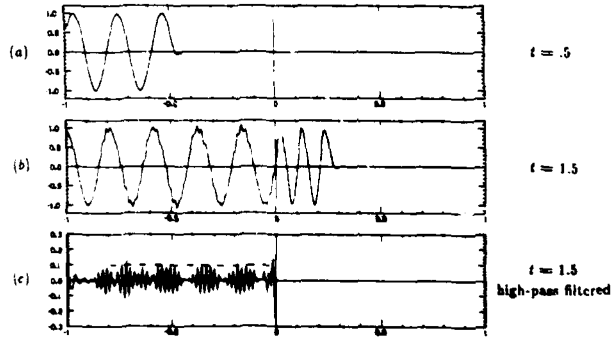


FIG. 3.4. Reflection and transmission at an LF4 interface. Same as Fig. 3.2 but with LF replaced by LF4.

values of z with $|z| = 1$ we expect one wavelike mode and one evanescent mode on each side, as shown in Fig. 3.3.

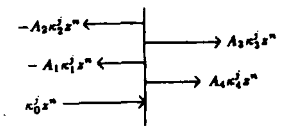


FIG. 3.3

Fig. 3.4 shows a repetition of Fig. 3.2 with LF replaced by LF4. Qualitatively, the behavior appears as before, except for one interesting change: the reflected parasite now travels at speed $C \approx -5/3$, not -1. This is in keeping with Fig. 1.1c and with (1.2.8) (or (1.5.3) for $\ell = 4$). Let us predict the amplitude of the reflected parasite. For the given problem $z \approx 1$, and so (2.1.12) implies

$$0 \approx z - \frac{1}{z} = \frac{4\lambda a_\pm}{3}\left(\kappa - \frac{1}{\kappa}\right) - \frac{\lambda a_\pm}{6}\left(\kappa^2 - \frac{1}{\kappa^2}\right),$$

i.e.

$$\kappa^4 - 8\kappa^3 + 8\kappa - 1 \approx 0$$

on both sides of the interface. The zeros of this polynomial are

$$\kappa = 1, \quad -1, \quad 4 - \sqrt{15}, \quad 4 + \sqrt{15}.$$

The first two values correspond to right- and leftgoing wave modes, respectively, and the second two to right- and leftgoing evanescent modes. We will order the κ_ν's according to

$$\kappa_0 \approx 1, \quad \kappa_1 \approx -1, \quad \kappa_2 \approx 4 + \sqrt{15}, \quad \kappa_3 \approx 1, \quad \kappa_4 \approx 4 - \sqrt{15},$$

but we will need a little more precision for κ_0 and κ_3, namely (as in Example 3.1)

$$\kappa_0 \approx e^{.30i} \approx 1 + .30i, \qquad \kappa_3 \approx e^{.60i} \approx 1 + .60i.$$

Now from (3.2.17) we obtain the amplitude sought,

$$\begin{aligned} A_1 &= \frac{(\kappa_2 - \kappa_0)(\kappa_3 - \kappa_0)(\kappa_4 - \kappa_0)\kappa_1^{3/2}}{(\kappa_2 - \kappa_1)(\kappa_3 - \kappa_1)(\kappa_4 - \kappa_1)\kappa_0^{3/2}} \\ &\approx \frac{(3 + \sqrt{15})(.30i)(3 - \sqrt{15})(-i)}{(5 + \sqrt{15})(2)(5 - \sqrt{15})(1)} = \frac{-1.8}{20} = -.09. \end{aligned}$$

An exact calculation from (3.2.17) gives the slightly larger result

$$A_1 \approx -.100476 + .001246i, \qquad |A_1| \approx .100484.$$

These numbers are in good agreement with the magnitude of the wiggles observed in Fig. 3.4b, which are once again isolated by a high-pass filter in Fig. 3.4c.

Example 3.4: Mesh-refinement interfaces

Instead of considering a discontinuous coefficient, let us now look at problems where the mesh size changes discontinuously at $x = 0$. We will stick to the equation $u_t = au_x$ and to models with one leftgoing and one rightgoing mode. Assume that a rightgoing signal $\kappa_r^j z^n$ hits the interface from the left, generating steady state reflected and transmitted signals $A\kappa_l^j z^n$ and $B\kappa_r^j z^n$. We will calculate A and B for three different kinds of mesh refinement.

(i) Crude mesh refinement. The reflection and transmission properties of the following mesh-refinement scheme (in its semi-discrete limit) are analysed by Vichnevetsky in [Vi81b]. Let x_j denote jh_- for $j \le 0$ and jh_+ for $j \ge 0$, where h_- and h_+ are arbitrary. Let (3.2.51) be modeled at all points $j \ne 0$ by LF, or more generally by any three-point linear multistep formula (2.4.2),

$$\frac{\rho(Z)}{k\sigma(Z)} v_j^n = a\frac{v_{j+1}^n - v_{j-1}^n}{2h_\pm}, \tag{3.2.18}$$

and at $j = 0$ by the related formula

$$\frac{\rho(Z)}{k\sigma(Z)} v_j^n = a\frac{v_1^n - v_{-1}^n}{h_- + h_+}, \tag{3.2.19}$$

as illustrated in Fig. 3.5.

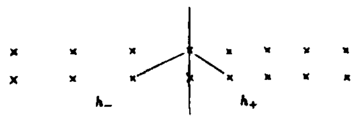


FIG. 3.5

The interface formulas for this model are then

$$1 + A = B,$$

$$\frac{\rho(Z)}{k\sigma(Z)}(1 + A) = \frac{a}{h_- + h_+}(B\kappa_r - 1/\kappa_r - A/\kappa_l).$$

After making use of the first formula, the second can be rewritten

$$\frac{h_-\rho(z)}{k\sigma(z)} + \frac{h_-A\rho(z)}{k\sigma(z)} + \frac{h_+\rho(z)}{k\sigma(z)}(1+A) = a(1+A)\kappa_t - a/\kappa_i - aA/\kappa_r.$$

The quantities $z, k, h_\pm, a$ can be eliminated from this equation by means of (3.2.18) or (2.4.3), and one obtains

$$\frac{1}{2}(\kappa_i - 1/\kappa_i) + \frac{1}{2}A(\kappa_r - 1/\kappa_r) + \frac{1+A}{2}(\kappa_t - 1/\kappa_t) = (1+A)\kappa_t - 1/\kappa_i - A/\kappa_r,$$

hence

$$(\kappa_i + 1/\kappa_i) + A(\kappa_r + 1/\kappa_r) - (1+A)(\kappa_t + 1/\kappa_t) = 0,$$

which implies

$$A = \frac{(\kappa_t + 1/\kappa_t) - (\kappa_i + 1/\kappa_i)}{(\kappa_r + 1/\kappa_r) - (\kappa_t + 1/\kappa_t)}.$$

By (2.4.8), this leads to

$$A = \frac{(\kappa_i + 1/\kappa_i) - (\kappa_t + 1/\kappa_t)}{(\kappa_i + 1/\kappa_i) + (\kappa_t + 1/\kappa_t)}, \qquad B = \frac{2(\kappa_i + 1/\kappa_i)}{(\kappa_i + 1/\kappa_i) + (\kappa_t + 1/\kappa_t)}. \tag{3.2.20}$$

An alternative expression for these results is

$$A = \frac{\cos\xi_i h - \cos\xi_t h}{\cos\xi_i h + \cos\xi_t h}, \qquad B = \frac{2\cos\xi_i h}{\cos\xi_i h + \cos\xi_t h}. \tag{3.2.21}$$

Compare [Vi81b], eq. (26).

(ii) Coarse mesh approximation. Suppose that in the above setup, h_+ is an integral multiple of h_-: $h_+ = mh_-$. Then instead of (3.2.19), it is natural to consider applying the coarse mesh formula used for $j \geq 0$ at $j = 0$ also, with the lefthand value needed taken from $j = -m$, as illustrated in Fig. 3.6:

$$\frac{2\rho(Z)}{\sigma(Z)}v_0^n = a\lambda_+(v_1^n - v_{-m}^n). \tag{3.2.22}$$

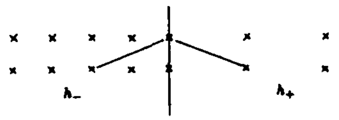


FIG. 3.6

With this interface condition in effect, the interface formulas become

$$1 + A = B,$$

$$\frac{2\rho(z)}{\sigma(z)}(1+A) = a\lambda_+(\kappa_t B - \kappa_i^{-m} - A\kappa_r^{-m}),$$

which by means of (3.2.18) reduce to

$$a\lambda_+(\kappa_t - 1/\kappa_t)(1+A) = a\lambda_+\left(\kappa_t(1+A) - \kappa_i^{-m} - A\kappa_r^{-m}\right),$$

i.e.

$$(1+A)/\kappa_t = \kappa_i^{-m} + A\kappa_r^{-m},$$

and therefore

$$A = \frac{\kappa_i^{-m} - 1/\kappa_t}{1/\kappa_t - \kappa_r^{-m}}. \tag{3.2.23}$$

By (2.4.8), this leads to

$$A = \frac{\kappa_i^{-m} - 1/\kappa_t}{1/\kappa_t - (-\kappa_i)^m}, \qquad B = \frac{\kappa_i^{-m} - (-\kappa_i)^m}{1/\kappa_t - (-\kappa_i)^m}. \tag{3.2.24}$$

(iii) BKO mesh refinement. The following "BKO" scheme was proposed by Browning, Kreiss, and Oliger in [Br73], and some of its reflection properties are analyzed in [Vi81b]. Suppose again that h_- and h_+ are arbitrary. Now let the left- and righthand grids overlap, as follows:

right: $x_j^+ = jh_+ \qquad j = -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots,$

left: $x_j^- = jh_- \qquad j = \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}, \ldots.$

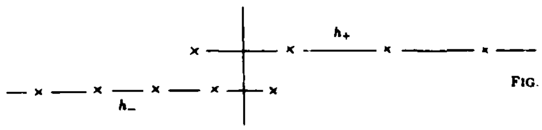


FIG. 3.7

Then, as a discrete analog of C^1 continuity at $x = 0$, consider the interface conditions

$$\begin{gathered} v_{\frac{1}{2}}^- + v_{-\frac{1}{2}}^- = v_{\frac{1}{2}}^+ + v_{-\frac{1}{2}}^+, \\ \frac{1}{h_-}(v_{\frac{1}{2}}^- - v_{-\frac{1}{2}}^-) = \frac{1}{h_+}(v_{\frac{1}{2}}^+ - v_{-\frac{1}{2}}^+) \end{gathered} \tag{3.2.25}$$

(with obvious notation). The corresponding interface formulas are

$$\kappa_i^{\frac{1}{2}} + \kappa_i^{-\frac{1}{2}} + A(\kappa_r^{\frac{1}{2}} + \kappa_r^{-\frac{1}{2}}) = B(\kappa_t^{\frac{1}{2}} + \kappa_t^{-\frac{1}{2}}),$$

$$\frac{1}{h_-}(\kappa_i^{\frac{1}{2}} - \kappa_i^{-\frac{1}{2}}) + \frac{A}{h_-}(\kappa_r^{\frac{1}{2}} - \kappa_r^{-\frac{1}{2}}) = \frac{B}{h_+}(\kappa_t^{\frac{1}{2}} - \kappa_t^{-\frac{1}{2}}).$$

These have the solution

$$A = \frac{h_- \cot\theta_i - h_+ \cot\theta_t}{h_+ \cot\theta_t - h_- \cot\theta_r}, \qquad B = \frac{h_- \cot\theta_i - h_- \cot\theta_r}{h_+ \cot\theta_t - h_- \cot\theta_r}, \tag{3.2.26}$$

where $\theta_i = \xi_i h/2$ again, so that $\cot\theta_i = i(\kappa_i^{\frac12} + \kappa_i^{-\frac12})/(\kappa_i^{\frac12} - \kappa_i^{-\frac12})$, and so on. For the case of three-point linear multistep formulas, (3.2.7) converts the result to

$$A = \frac{h_- \cot\theta_i - h_+ \cot\theta_t}{h_+ \cot\theta_t - h_- \tan\theta_i}, \qquad B = \frac{h_- \cot\theta_i - h_- \tan\theta_i}{h_+ \cot\theta_t - h_- \tan\theta_i}. \tag{3.2.27}$$

For LF^2, similarly, (3.2.12) reduces (3.2.26) to

$$A = \frac{h_- \cot\theta_i - h_+ \cot\theta_t}{h_+ \cot\theta_t + h_- \cot\theta_i}, \qquad B = \frac{2h_- \cot\theta_i}{h_+ \cot\theta_t + h_- \cot\theta_i}. \tag{3.2.28}$$

Again, these equations are similar to the admittance formulas (3.2.13).

Example 3.5: Boundaries

Finally, to justify the title of this chapter we must consider some problems containing boundaries rather than interfaces, at which there will be reflected but not transmitted signals. Let the equation

$$u_t = a u_x \qquad x \geq 0, \quad a \neq 0$$

be modeled by a difference formula Q on the grid $x_j = jh$, $j = 0, 1, 2, \ldots$. If Q extends ℓ points to the left of center, then numerical boundary formulas will be needed for the points $j = 0, \ldots, \ell-1$. Let us assume $\ell = 1$, so that only one boundary formula is needed. If a strictly leftgoing signal $\kappa_i^j z^n$ hits the point $j = 0$, then in the steady state some energy will propagate rightward as a signal $A\kappa_r^j z^n$. We seek the reflection coefficient A.

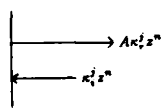


FIG. 3.8

Suppose first that the boundary formula is $(q-1)$**st-order space extrapolation,**

$$S: \qquad (K-1)^q v_0^{n+1} = 0, \qquad q \geq 1 \tag{3.2.29}$$

where K is the spatial shift operator defined in §2.1. Then A will satisfy

$$(\kappa_i - 1)^q + A(\kappa_r - 1)^q = 0,$$

hence

$$A = -\left(\frac{\kappa_i - 1}{\kappa_r - 1}\right)^q. \tag{3.2.30}$$

If Q is LF or any other three-point linear multistep formula, then (2.4.8) converts this to

$$A = -\left(\frac{\kappa_i - 1}{-1/\kappa_i - 1}\right)^q = -\left(\frac{1-\kappa_i}{1+\kappa_i}\right)^q \kappa_i^q. \tag{3.2.31}$$

Suppose alternatively that the boundary formula is $(q-1)$**st-order space-time extrapolation,**

$$ST: \qquad (KZ^{-1} - 1)^q v_0^n = 0, \qquad q \geq 1. \tag{3.2.32}$$

Now A will satisfy

$$(\kappa_i - z)^q + A(\kappa_r - z)^q = 0,$$

hence

$$A = -\left(\frac{\kappa_i - z}{\kappa_r - z}\right)^q. \tag{3.2.33}$$

For three-point linear multistep schemes this becomes

$$A = -\left(\frac{z - \kappa_i}{1 + z\kappa_i}\right)^q \kappa_i^q. \tag{3.2.34}$$

This is an example in which it is not practical to eliminate z from the formula.

• • •

Our purpose in this section has been to show how reflection and transmission coefficients can be computed, not to apply such computations to the evaluation of particular numerical treatments of boundaries or interfaces. But obviously this kind of information is potentially useful if one is trying to choose between various numerical methods.

The reflection and transmission behavior we have predicted, like the group velocity phenomena of Chapter 1, can be readily confirmed with numerical experiments. We have performed a number of these, but except for Demos. 3.1 and 3.2 already presented, we will not take the space to describe them here.

3.3 Energy flux and energy conservation*

Suppose that a Cauchy stable difference formula admits the wave solution

$$v_j^n = A\kappa^j z^n \qquad (3.3.1)$$

for some constant A, with $|z| = 1$. If $|\kappa| = 1$, then by Thm. 2.3.1, the wave has a well defined group velocity $C \in \mathbb{R}$. It is natural to define the **energy flux** (magnitude) Φ of (3.3.1) as the absolute group speed times the square of the amplitude,

$$\Phi = |A|^2|C| \qquad \text{if } |z| = |\kappa| = 1. \qquad (3.3.2)$$

One might prove that asymptotically, Φ measures the ℓ_2 energy flow per unit time across a given line $x = x_0$. Other definitions could be used for energies other than ℓ_2. If $|z| = 1$ and $|\kappa| \neq 1$, then there is no energy flux,

$$\Phi = 0 \qquad \text{for } |z| = 1,\ |\kappa| \neq 1. \qquad (3.3.3)$$

If $|z| > 1$, in which case $|\kappa| \neq 1$ by Thm. 2.2.1, then a sensible definition of Φ would have to vary with x and increase with t. But we will not define Φ in this case.

Given a difference model containing a boundary or interface, we naturally ask: does the interface conserve energy? If not, how close does it come? For the steady state solutions of the last section, we have all the machinery in place to answer these questions. Assume, for example, that $u_t = au_x$ is modeled by one three-point difference scheme for $x < 0$ and another for $x > 0$, and that a rightgoing wave (3.3.1) is incident on the interface at $x = 0$. In the steady state, reflected and transmitted waves will be generated. We define the **efficiency** of the interface for the given wave, E, by the formula

$$E = \frac{\Phi_r + \Phi_t}{\Phi_i}. \qquad (3.3.4)$$

Energy is absorbed, conserved, or created at the interface if $E < 1$, $E = 1$, or $E > 1$, respectively. More generally, if an interface generates a collection of outgoing signals in response to a collection of incoming ones, then the efficiency for that configuration is

$$E = \sum_{\text{outgoing}} \Phi_\nu \Big/ \sum_{\text{incoming}} \Phi_\nu \qquad (3.3.5)$$

*Many of the ideas in this section appear in [Vi81b].

Of the mesh-refinement problems considered in Example 3.5 of the last section, two conserve energy exactly: "crude mesh refinement" for any three-point linear multistep formula Q, and BKO mesh refinement for LF^2. Let us verify these claims. For Q applied to $u_t = au_x$ we have by (2.4.7),

$$C = \cos\xi h\, f(z)$$

for some function f. From this formula, the reflection and transmission coefficients (3.2.21) for crude mesh refinement become

$$A = \frac{C_i - C_t}{C_i + C_t}, \qquad B = \frac{2C_i}{C_i + C_t}. \qquad (3.3.6)$$

Inserting these values in (3.3.2) now gives the fluxes [cf. [Cl76], eq. (8-1-6)]

$$\Phi_i = C_i, \qquad \Phi_r = C_i\frac{(C_i - C_t)^2}{(C_i + C_t)^2}, \qquad \Phi_t = \frac{4C_i^2 C_t}{(C_i + C_t)^2},$$

and applying these to (3.3.4) yields

$$E = \frac{(C_i - C_t)^2 + 4C_iC_t}{(C_i + C_t)^2} = 1,$$

as claimed. Similarly, for the case of LF^2 with BKO mesh refinement, eqs. (1.8.7) and (1.8.8) (ignoring the terms in η) imply

$$C_i = \frac{\lambda_- \sin\xi_i h}{\sin\omega k} = \frac{2\lambda_- \sin\theta_i \cos\theta_i}{2\sin\frac{\omega k}{2}\cos\frac{\omega k}{2}}$$

$$= \frac{\lambda_- \sin\theta_i\cos\theta_i}{\sin\frac{\omega k}{2}\cos\frac{\omega k}{2}}\left(\frac{\sin^2\frac{\omega k}{2}}{\lambda_-^2\sin^2\theta_i}\right) = \frac{h_- \cot\theta_i}{k\cot\frac{\omega k}{2}},$$

with corresponding expressions for C_r and C_t. From this formula and (3.2.28), it follows that (3.3.6) holds for this problem too, and this implies $E = 1$ as before.

However, it is only in exceptional cases that a boundary or interface conserves energy exactly. The reason is that for this to happen, the errors introduced by the interior formulas and the interface formulas must exactly counterbalance, so the two sets of formulas must be fortuitously compatible in some sense. In particular, the other mesh refinement problems of the last section, such as LF with BKO or with the coarse mesh approximation, do not exactly conserve energy.

Energy conservation is an attractive property, especially if extensions to nonlinear problems are being considered, but one should not automatically assume that if one interface exactly conserves energy and another does not, then the former is better. For LF applied to $u_t = au_x$, for example, (3.2.27) implies that the nonconserving BKO interface generates a reflected parasite of amplitude $A = O(h^3)$, while (3.2.21) gives $A = O(h^2)$ for the "crude" interface. Surely it is no virtue of the latter scheme that the spurious signal it generates on the left is large enough to balance the flux error it introduces on the right.

3.4 Cutoff frequencies and evanescent waves

We have observed earlier that although a nondissipative difference model Q must admit waves of all wave numbers $\xi \in [-\pi/h, \pi/h]$, the same is not true for all frequencies $\omega \in [-\pi/k, \pi/k]$. A frequency that corresponds to no wave solutions may be said to lie in the **stop band** or **forbidden band** for Q. In Figs. 1a-c, these are the values of ω for which no value of ξ appears on the plot. Of course there will be some wave number ξ for every ω, since Q must do *something* in response to a forcing oscillation $\sin \omega t$, but for ω in the stop band ξ will be complex, corresponding to an evanescent mode that by (3.3.3) carries no energy.

In a problem involving an interface, it may happen that a frequency for which a wave may exist on one side lies in the stop band on the other. In this event the response to such an incident wave will be $\phi_t = 0$ — total reflection. Given an interface, one may look for the minimum frequency ω_c, the **cutoff frequency**, at which a transmitted wave cannot exist. The solution to this problem will satisfy the cutoff condition

$$C(\xi_c, \omega_c) = 0. \tag{3.4.1}$$

One can see this by considering that in a dispersion plot such as Fig. 1.1, ω_c is associated with a zero slope. Algebraically, the explanation is that if $z(\kappa)$ has $|z| = 1$ for $\arg\kappa \le \arg\kappa_c$ and $|z| \ne 1$ for $\arg\kappa > \arg\kappa_c$, then κ_c must be a multiple root, which we know by Thm. 2.3.1 corresponds to $C = 0$.

Cutoff frequencies for finite difference and finite element models have been discussed previously in [Ok76], [Br73], and [Vi80].

The primary significance of (3.4.1) is that it enables one to determine cutoff frequencies by solving an algebraic equation. Another interesting implication is that although the vanishing of the transmitted wave as ω rises above ω_c may be discontinuous (the transmitted wave abruptly becomes evanescent, but its amplitude

does not become zero), the vanishing of the transmitted energy flux is not. Instead, Φ decreases smoothly to 0 as $\omega \uparrow \omega_c$, since C decreases smoothly to 0.

Vichnevetsky points out that in the case of interfaces between LF models, the evanescent wave that appears for $\omega > \omega_c$ always has wavelength $4h$ [Vi80,Vi81b]. The explanation is that by (2.1.8), $|z| = 1$ and $|\kappa| \ne 1$ can only happen with κ pure imaginary, which amounts to wavelength $4h$. The "$4h$ phenomenon" does not extend to arbitrary difference models, however.

3.5 Reflection of a general wave packet

By the methods described so far we can now determine exactly how a monochromatic signal $e^{i(\omega t - \xi x)}$ is reflected and transmitted at a boundary or interface. The question is, *how can this information be used to predict the reflection and transmission of a general wave packet?* The problem is one of Fourier synthesis in an inhomogeneous medium, and it is subtle.* One might expect, for example, that if the reflection coefficients satisfy $|A(\xi)| \le A_{\max} < \infty$ for all ξ, then a general estimate $\|v^n\|_2 \le A_{\max}\|v^0\|_2$ will hold. However, Thm. 4.2.3 will show that this is not the case.

We will study the simplest possible example. Let the equation

$$u_t = u_x, \qquad x, t \ge 0$$

be modeled by a finite difference scheme $\tilde{Q}$ on the grid $(x_j, t_n) = (jh, nk)$ for $j, n \ge 0$, with $h = 1$ for convenience. Let $\tilde{Q}$ consist of $Q =$ CN (1.1.16) for all points $j, n \ge 1$ coupled with some two-level boundary equation for $j = 0$, $n \ge 1$. For initial data we take

$$u_j^0 = f_j, \qquad j \ge 0$$

for some sequence f. Now v^{n+1} is completely determined by v^n. Since CN is nondissipative and x-reversing, we expect significant reflections at the boundary.

Let ℓ_2^+ denote the set of square-summable sequences $(f_j)_{j\ge 0}$. If $f \in \ell_2^+$, then it has a Fourier representation

$$f_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-i\xi j}\hat{f}(\xi)\,d\xi \tag{3.5.1}$$

*A solution for a special case of this problem is sketched in §6 of [Vi81b], but it appears to be invalid except, perhaps, in some asymptotic sense. For example, this solution begins by considering a wave packet with compact support whose transform also has compact support, and such a combination cannot occur.

for some function $\hat{f} \in L_2[-\pi,\pi]$, and $\hat{f}$ is given by the Fourier transform

$$\hat{f}(\xi) = \sum_{j=0}^{\infty} f_j e^{i\xi j}. \tag{3.5.2}$$

By (1.1.18) and (1.2.7), we know that for each $\xi_\ell \in [-\frac{\pi}{2},\frac{\pi}{2}]$, CN admits a leftgoing wave, and at the same frequency there is one corresponding rightgoing wave with wave number $\xi_r = \pi - \xi_\ell$ (Fig. 1.1c). (Here ξ_r should be taken modulo 2π.) Let $A(\xi)$ denote the corresponding reflection coefficient function for monochromatic solutions for the given boundary conditions. Now suppose that by chance $\hat{f}$ happens to satisfy the reflection condition

$$\hat{f}(\pi-\xi) = A(\xi)\hat{f}(\xi) \qquad \text{for } \xi \in [-\tfrac{\pi}{2},\tfrac{\pi}{2}]. \tag{3.5.3}$$

Then by the definition of $A(\xi)$, f is the superposition of steady state solutions of $\hat{Q}$:

$$f_j = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2}\left[e^{-i\xi j} + A(\xi)e^{-i(\pi-\xi)j}\right]\hat{f}(\xi)\,d\xi. \tag{3.5.4}$$

Therefore if $\{v^n\}$ is computed with f as initial data, then each steady-state solution evolves under $\hat{Q}$ in a uniform fashion, oscillating according to a factor $e^{i\omega(\xi)t}$, and we obtain a Fourier representation for v_j^n valid for all n:

$$v_j^n = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} e^{i\omega(\xi)t}\left[e^{-i\xi j} + A(\xi)e^{-i(\pi-\xi)j}\right]\hat{f}(\xi)\,d\xi \qquad (t = nk). \tag{3.5.5}$$

In general, of course, $\hat{f}$ will not satisfy (3.5.3). The main idea of this section is as follows. Consider choosing arbitrary values f_j for $j < 0$ so that f is extended to a biinfinite sequence $(f)_{j\in \mathbb{Z}} \in \ell_2$. Any such sequence will have a Fourier representation (3.5.1), where now $\hat{f} \in L_2[-\pi,\pi]$ is given by

$$\hat{f}(\xi) = \sum_{j=-\infty}^{\infty} f_j e^{i\xi j}. \tag{3.5.6}$$

Suppose an extension f can be found for which (3.5.3) holds. Then again, (3.5.5) must give v_j^n for all n. In fact, (3.5.5) will determine a function $\{v_j^n\}$ that satisfies CN for *all* $j \in \mathbb{Z}$, and in addition satisfies the boundary equation imposed by $\hat{Q}$ at $j = 0$. Therefore its restriction to $j \geq 0$ must be exactly the solution we seek.

We can therefore determine the reflection in $j = 0$ of a general wave packet if we can solve the following problem:

REFLECTION PROBLEM

Given: (i) f_j for $j \geq 0$

(ii) $\hat{f}$ satisfies (3.5.3) for a known function $A(\xi)$

Find: (i) f_j for all $j \in \mathbb{Z}$

(ii) $\hat{f}(\xi)$ for all $\xi \in [-\pi,\pi]$

In effect (i) gives us half of f, and (ii) gives us half of its transform. The parameter count appears right for the problem to be well posed.

The reflection problem as stated has a simple interpretation. Given initial data $(f_j)_{j\geq 0}$, we seek a distribution of **dual initial data** $(f_j)_{j<0}$ such that as n increases, the solution v^n obtained by applying CN for all $j \in \mathbb{Z}$ satisfies the boundary equation of $\hat{Q}$ at $j = 0$. In other words, the dual packet must be chosen so that it contains rightgoing components that exactly duplicate any reflections of the initial data that should be observed under $\hat{Q}$. The idea is illustrated in Fig. 3.9:

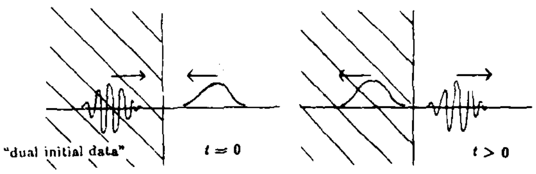


FIG. 3.9

Mathematically, the reflection problem amounts to the problem of solving an integral equation. Let f_+ and $\hat{f}_+$ denote the restrictions of f and $\hat{f}$ to $j \geq 0$ and $\xi \in [-\frac{\pi}{2},\frac{\pi}{2}]$, respectively. According to (3.5.4), we need to solve the equation

$$\Psi\hat{f}_+ = f_+, \tag{3.5.7}$$

for $\hat{f}_+$, where $\Psi : L_2[-\frac{\pi}{2},\frac{\pi}{2}] \to \ell_2^+$ denotes the integral operator

$$(\Psi\hat{f}_+)_j = \int_{-\pi/2}^{\pi/2} K(j,\xi)\hat{f}_+(\xi)\,d\xi, \tag{3.5.8}$$

where K denotes the kernel

$$K(j,\xi) = \frac{1}{2\pi}\left[e^{-i\xi j} + A(\xi)e^{-i(\pi-\xi)j}\right]. \tag{3.5.9}$$

Unfortunately, we have not yet made any progress in solving the problem as formulated here. It appears that it might be possible to treat the integral equation by some variant of the ***Wiener-Hopf technique*** [Mo53], which is designed to handle Fourier transforms that are split into two halves. However, the solution remains to be worked out. It seems that despite the obvious likelihood that there is a connection between this problem and the Wiener-Hopf methods of Strang and Osher mentioned in §0.2, the two formulations are not the same.

3.6 General formulation; the "folding trick"

We will now write down formally the linear algebraic relations that govern steady-state solution behavior for a system of equations at a boundary or interface. In doing so we face the question of how much flexibility to permit in the representation of a difference model. For example, in the interface calculations of §3.2, it was sometimes convenient to use a grid $j = 0, \pm 1, \pm 2, \ldots$, and sometimes $j = \pm\frac{1}{2}, \pm\frac{3}{2}, \ldots$ was better. At issue is a tradeoff between the simplicity of the general formulation and the simplicity of its application to particular problems. Our procedure will be to present the generalities in a restricted formalism here, but continue to abuse that formalism later as convenient for dealing with particular cases.

Our main simplification will be that instead of treating interface problems as interface problems, we will reduce them formally to boundary value problems by a device known as the **folding trick**. If the original problem is made up of a system in N_1 unknowns on the left coupled with a system in N_2 unknowns on the right, the folding trick consists of replacing these in the obvious way by an equivalent system in $N_1 + N_2$ unknowns involving only a boundary. This device has been used in various papers on numerical stability, including [Ci71,Ci72,Br73,Su74]. It is not an unqualified blessing, however, for it tends to obscure what is really going on when one deals with an interface. In particular, one must remember that the system obtained after folding is not an arbitrary system in $N_1 + N_2$ variables, but a 2×2 block diagonal system, since the left-side and right-side variables are uncoupled except through the boundary conditions. In particular, it follows that if the difference models on each side of an interface satisfy Assumption 2.1 (diagonalizability – see §2.5), then that assumption also holds for the folded problem.

Consider then the $(s+2)$-level N-vector difference model Q of (2.5.2). We assume

that Q satisfies Asssmption 2.1. In addition, corresponding to Assumption 3.1, let us impose the following condition:

Assumption 3.2. *For all z with $|z| \geq 1$, Q admits exactly n_ℓ leftgoing and n_r rightgoing solutions (2.5.8), where n_ℓ and n_r are some fixed integers.* //

Instead of letting j range over all integers, we now restrict it to $j \geq 0$. Q will apply at all points $j \geq \ell$, and n_r additional boundary conditions are then in general needed that involve v_j^{n+1}, $j = 0, \ldots, \ell - 1$. We can write these in the form

$$\sum_{j=0}^{j_{\max}} \sum_{\sigma=-1}^{\sigma_{\max}} S_{j\sigma} v_j^{n-\sigma} = 0 \tag{3.6.1}$$

for some integers $j_{\max}, \sigma_{\max} < \infty$, where each $S_{j\sigma}$ is a constant $n_r \times N$ matrix. The "0" on the right denotes the null vector of length n_r. We let $\bar{Q}$ denote the difference model consisting of Q for $j \geq \ell$ combined with (3.6.1). For $\bar{Q}$ to be usable we need a solvability assumption (cf. Ass. 3.1 of [Gu72] and Ass. 1.1 of [Mi81]).

Assumption 3.3. The model $\bar{Q}$ can be solved boundedly in the sense that if $v^{n-\sigma_{\max}}, \ldots, v^n \in \ell_2$ are given, then v^{n+1} is uniquely determined, and it satisfies a bound

$$\|v^{n+1}\|_2 \leq M \sum_{\sigma=0}^{\sigma_{\max}} \|v^{n-\sigma}\|_2.$$

The two-norm here is defined (cf. (2.2.1)) by

$$\|\phi\|_2^2 = h \sum_{j=0}^{\infty} |\phi_j|^2, \tag{3.6.2}$$

where $|\phi_j|$ denotes the vector two-norm. //

It can be shown that such a solvability assumption can hold only when (3.6.1) has n_r rows, as we have assumed (cf. [Mi81], Thm. 1.1).

Let z be a complex constant satisfying $|z| \geq 1$. According to (2.5.12), the general solution to (2.5.2) can be written

$$v_j^n = \sum_{i=1}^{n_r} a_i \kappa_i^j j^{k_i} \psi_i + \sum_{i=n_r+1}^{n_r+n_\ell} a_i \kappa_i^j j^{k_i} \psi_i \tag{3.6.3}$$

for some constants $\{a_i\}$. The two sums represent rightgoing and leftgoing signals, respectively. Let this formula be inserted in (3.6.1). The result is a linear system of equations of dimension $n_r \times N$ that involves $\{a_i\}$, $\{S_{j\sigma}\}$, [illegible], $\{\psi_i\}$, and z.

All of these quantities are known, since z is given, except for $\{a_i\}$. We can therefore rewrite (3.6.1) as the new system

$$\sum_{i=1}^{n_r} [\cdots] a_i + \sum_{i=n_r+1}^{n_r+n_\ell} [\cdots] a_i = 0,$$

where each term in brackets is an n_r-vector depending on z. If we write now

$$a^{[r]} = (a_1, \ldots, a_{n_r})^T, \qquad a^{[\ell]} = (a_{n_r+1}, \ldots, a_{n_r+n_\ell})^T,$$

then these equations take the form (cf. eq. (10.2) of [Gu72])

$$D^{[r]}(z) a^{[r]}(z) + D^{[\ell]}(z) a^{[\ell]}(z) = 0, \tag{3.6.4}$$

where $D^{[r]}$ is $n_r \times n_r$ and $D^{[\ell]}$ is $n_r \times n_\ell$. This equation represents the interface formulas for the general problem (2.5.2), (3.6.1).

Now we can solve for reflection coefficients. In the previous sections we just studied the repsonse to a single incident signal, but of course linearity implies that the response to a sum of incident signals will be the sum of the responses to each. The general problem of finding reflection coefficients is therefore: *given $a^{[\ell]}$, find $a^{[r]}$*. If $D^{[r]}$ is invertible, then (3.6.4) gives the result

$$a^{[r]} = -(D^{[r]})^{-1} D^{[\ell]} a^{[\ell]}. \tag{3.6.5}$$

This equation is the general solution to the problem of finding reflection coefficients, and $(D^{[r]})^{-1} D^{[\ell]}$ is an $n_r \times n_\ell$ matrix that might be called the **reflection coefficient matrix.** If the problem (2.5.2), (3.6.1) came from an interface problem by folding, then $a^{[\ell]}$ describes incident signals and $a^{[r]}$ both reflected and transmitted ones.

It is by *no means* always true that $D^{[r]}$ is invertible. In certain circumstances the examples of §3.2 demonstrate this problem. In Example 3.1, for instance, if $a_- > 0 > a_+$, then for $z = 1$ one has $\theta_r = \theta_\ell = 0$, and the denominators in (3.2.6) are 0. Similarly in Example 3.4(ii), for $a < 0$ and $z = 1$ one has $\kappa_r = \kappa_\ell = 1$, and if m is even, then the denominators in (3.2.24) are 0. Apparently for the wrong value of z in these problems, the reflection and transmission coefficients become infinite—and for nearby values, arbitrarily large. This is no flaw in our formulation, but the actual behavior of these schemes, as we will verify by experiment in the next chapter (Demo. 4.2). We will see there that the singularity of $D^{[r]}$ and the presence of infinite reflection coefficients are directly related to instability in finite difference models of

intitial boundary value problems.

• • •

All of our discussion in this chapter has been restricted to problems in one space dimension, but the same principles apply in the multidimensional case. Suppose for example that a plane wave with frequency ω and wave number vector $\xi^{(i)} = (\xi_1, \ldots, \xi_d)^T$ is incident upon a plane interface at $x_1 = 0$. As in the one-dimensional problem, the first step is to solve the dispersion relation to determine all possible reflected and transmitted wave number vectors $\xi^{(r)}, \xi^{(t)}$. Since the interface is parallel to the axes $x_2, \ldots, x_d$, one can use the fact that all components will have equal values of ω and $\xi_2, \ldots, \xi_d$, differing only in ξ_1. The various solutions ξ_1 on each side of the interface then yield waves oriented at various angles. The radiation condition requires that one pick out those waves with vector group velocities pointing away from the interface. Once these are determined, reflection and transmission coefficients can be computed as usual, and they will be angle-dependent. Thus ***Snell's Law for difference models***, already mentioned in §1.6, fits directly into the framework we have established for interface problems.

4. STABILITY FOR INITIAL BOUNDARY VALUE PROBLEMS

4.1 An example

From here on, the rest of the dissertation is concerned with the stability of finite difference models that contain boundaries or interfaces. According to the folding trick (§3.6), it is enough to consider the stability of models of initial boundary value problems. The available theory for this was developed by Kreiss, Osher, Gustafsson, and others in the decade preceding 1972, and was reported in an important paper by Gustafsson, Kreiss, and Sundström ("GKS") in 1972 [Gu72] (see §0.2 for further references). For further developments of this theory see [Gu75] and [Mi81], and for a systematic introduction to it see [Co80]. Our purpose in this chapter is to show that the key factor determining stability is dispersive wave propagation. We will see that the results of Kreiss and others are built around a group velocity test in a disguised form.

We will bring out our basic ideas with a simple example. Let the problem

$$u_t = u_x, \qquad u(x,0) = f(x) \tag{4.1.1}$$

be given on $x \geq 0$, $t \geq 0$; no boundary data at $x = 0$ are needed to make (4.1.1) well posed. To obtain an approximate solution on the grid $j, n \geq 0$, we can specify initial values v_j^0 and v_j^1 for $j \geq 0$, and apply LF (1.1.6) for $n \geq 2$ at points $j \geq 1$. *An additional boundary formula is then needed for* v_0^n, $n \geq 2$. *Let us pick the* zeroth-order space extrapolation formula (3.2.29),

$$v_0^{n+1} = v_1^{n+1} \qquad (n \geq 1), \tag{4.1.2}$$

and proceed to step forward in time.

Instability as spontaneous radiation from the boundary

Instability refers to the unbounded amplification of small perturbations. Now imagine that at some pair of adjacent time steps a rounding error or other perturbation happens to be introduced that has the form of a wave front with $(\kappa, z) = (1, -1)$,

$$v_j^n = \begin{cases} (-1)^n & (jh \leq \epsilon) \\ 0 & (jh > \epsilon) \end{cases} \tag{4.1.3}$$

for some $\epsilon \gg h$. To be a little more careful, we could make v decrease smoothly to 0 near $x = \epsilon$ rather than abruptly. Then what will happen as t increases? At $j = 0$, (4.1.3) satisfies both LF and (4.1.2), so the oscillation (4.1.3) will persist. At $jh = \epsilon$, the wave front will move at the group speed for the given pair (κ, z) — which by (1.2.5) is +1. Thus as t increases the wave will propagate rightwards into $x \geq 0$ at speed 1. The initial perturbation, with sum-of-squares energy on the order of ϵ, will give rise to a growing solution with energy on the order of $\epsilon + t$. Since ϵ might be arbitrarily small (so long as h is decreased accordingly), this amounts to an *amplification* of the initial perturbation by an unbounded factor. *The difference scheme is unstable, because there exists a rightgoing wave that satisfies both the interior formula LF and the boundary condition (4.1.2).*

DEMONSTRATION 4.1. Of course few random perturbations look exactly like (4.1.3), but instability comes about because almost *any* data will excite this mode to some extent. One can verify this experimentally. Fig. 4.1 shows a computation on a grid with $h = 1/200$, $\lambda = 1/2$. For initial data we took $v_j^0 = v_j^1 = 0$ for all j except for the "random" nonzero initial values

$$v_1^0 = 1, \qquad v_0^1 = \frac{1}{2}, \qquad v_1^1 = \frac{1}{3}. \tag{4.1.4}$$

Figs. 4.1a-c show the resulting solution at steps $n = 1, 100, 200, 201$, i.e. $t = .0025$, .25, .5, .5025. Obviously the expected incoming mode has been excited, and apparently no others.

In a realistic computation, truncation errors would usually cause a similar radiation of energy in this mode from the boundary. From (1.2.5) or Fig. 1.1a we know that there are many other rightgoing modes for LF — in fact, any wave with $\xi h \leq \pi/2$ and $\omega k \geq \pi/2$ or $\xi h \geq \pi/2$ and $\omega k \leq \pi/2$. The mode $(\kappa, z) = (-1, 1)$ is the simplest example. None of these lead to instability, however, because none of them satisfy (4.1.2).

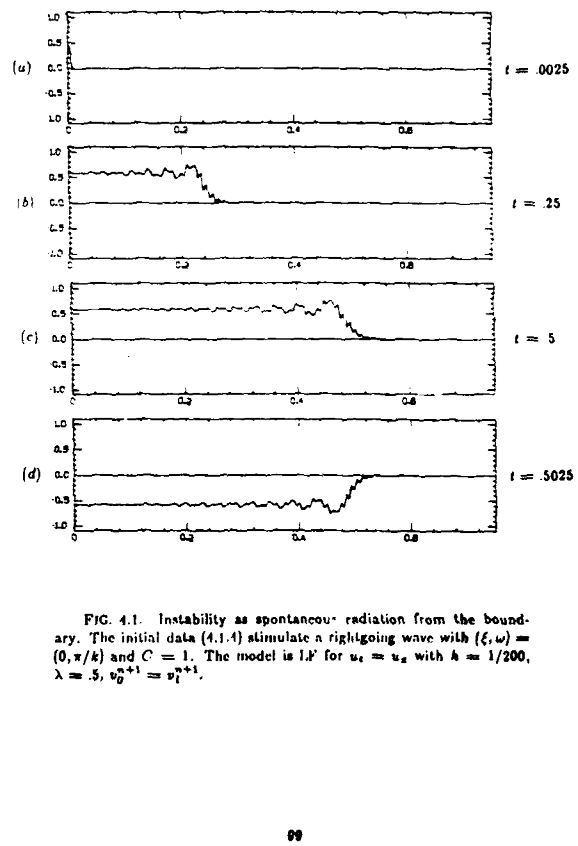


FIG. 4.1. Instability as spontaneous radiation from the boundary. The initial data (4.1.4) stimulate a rightgoing wave with $(\xi,\omega) = (0,\pi/k)$ and $C = 1$. The model is LF for $u_t = u_x$ with $h = 1/200$, $\lambda = .5$, $v_0^{n+1} = v_1^{n+1}$.

Instability as an infinite reflection coefficient

Another way to look at the instability of initial boundary value problems is in terms of reflection coefficients. In Example 3.5 we have considered the boundary condition (4.1.2) already, and derived the reflection coefficient formula (3.2.31)

$$A(z) = -\kappa_\ell\left(\frac{1-\kappa_\ell}{1+\kappa_\ell}\right), \qquad (4.1.5)$$

where $\kappa_\ell = \kappa_\ell(z)$ is the spatial variation factor for the incident leftgoing signal. From this formula it is evident that A becomes infinite if (and only if) $\kappa_\ell = -1$. By (2.1.8), LF for (4.1.1) has two modes with $\kappa = -1$, namely $(\kappa, z) = (-1, 1)$ and $(-1,-1)$. Of these the latter is the leftgoing one, and by (2.4.8), the corresponding reflected rightgoing mode is $(\kappa_r, z) = (1, -1)$. This is exactly the unstable mode we have identified in (4.1.3). *The difference scheme is unstable, because there exists a leftgoing wave for which the reflection coefficient is infinite.*

DEMONSTRATION 4.2. It is not possible to observe infinite amplification in reflection, but we can come arbitrarily close. Fig. 4.2 shows an experiment involving the same model as Demo. 4.1. In Fig. 4.2a, an initial Gaussian packet

$$v_j^1 = -v_j^0 = \frac{1}{10}(-1)^j e^{-[(jh-.2)/.025]^2} \qquad (4.1.6)$$

is shown for $t = n = 0$. As t increases, this packet moves left at speed $C(-1,-1) = -1$, hits the boundary, and reflects rightward. Fig. 4.2b shows the result at $t = 0.5$. One sees immediately that the reflected wave is not a packet, but a plane wave as in Fig. 4.1—the unstable mode has become lodged in the boundary, where it will continue to radiate forever. In addition, there has been an 18-fold amplitude increase from 0.1 to 1.7725.

By doubling the width of the initial packet, one doubles the reflected amplitude. Figs. 4.2c–d show the experiment repeated with the width .025 of (4.1.6) replaced by .05. Now the reflected amplitude is 3.5449—just twice the previous value. One can account for this in various ways. The simplest is to argue that the broadened pulse interacts with the boundary for twice as long, enabling twice as much of the unstable mode to accumulate there. A more elegant explanation starts from the fact that any finite packet cannot consist of energy at exactly the critical wave number $\xi_0 = 0$ (the uncertainty principle again), but will approximate ξ_0 with some effective wave number $\xi_{\rm eff}$. Eq. (4.1.5) suggests that the observed reflected amplitude should behave like

$$\text{amplitude} \approx \frac{\text{const.}}{|\xi_{\rm eff} - \xi_0|}. \qquad (4.1.7)$$

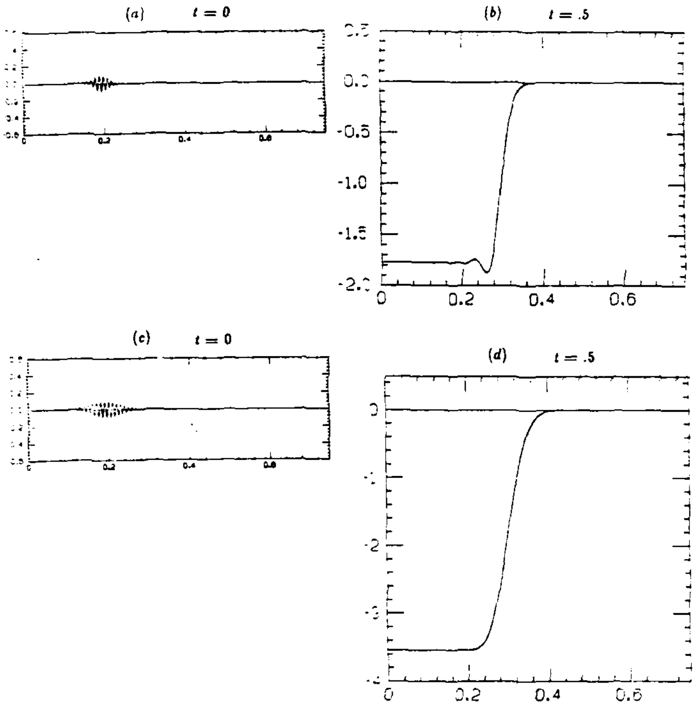


FIG. 4.2. Instability as an infinite reflection coefficient. The initial leftgoing wave packet (4.1.6) with $(\xi,\omega) = (\pi/h, \pi/k)$ hits the boundary and reflects as a wave front with $(\xi,\omega) = (0, \pi/k)$ of much greater amplitude. Doubling the width of the packet doubles the amplification. The model is LF for $u_t = u_x$ with $h = 1/200$, $\lambda = .5$, $v_0^{n+1} = v_1^{n+1}$.

By doubling the width of the packet, we have cut $\xi_0 - \xi_{\bullet\pi}$ in half, and thereby doubled (4.1.7). In §6.5 we will pursue this kind of reasoning in some detail. It is likely that by an extension of the ideas of §3.5, one could also get an exact expression for the reflected amplitude.

Instead of widening the packet, we could have made h smaller. As a general rule one can expect amplitude increases comparable to the number of grid points in the initial packet. For fine enough meshes this implies arbitrarily great increases in amplitude. This amounts to instability in any norm.

Discussion

Of course not all numerical boundary conditions are unstable. To obtain stability in the present problem, we might replace (4.1.2) by the zeroth-order space-time extrapolation formula (3.2.32),

$$v_0^{n+1} = v_1^n \qquad (n \geq 1). \tag{4.1.8}$$

From the corresponding equation $z = \kappa$ and the equation (2.1.8) for LF, it is immediate that now $\hat{Q}$ admits no regular solutions except $(\kappa, z) = (1, 1)$ or $(-1, -1)$. Since both of these are leftgoing, no spontaneous radiation from the boundary is possible. Similarly for the reflection coefficient point of view, (3.2.34) shows that $A = \infty$ is possible only for $z = -1/\kappa_\ell$, a condition that is never satisfied under LF.

Obviously the possibility of spontaneous rightgoing modes and the existence of infinite reflection coefficients are algebraically related, so our two approaches to instability are far from independent. They are however not equivalent, for it turns out that there are a number of problems that admit a spontaneous rightgoing mode, but for which all reflection coefficients are uniformly bounded. To what extent such models act unstable in practice is open to question, and these are among the "borderline cases" of stability to be discussed in §5. Chapter 5 is also concerned with another weakly unstable borderline case, namely the situation in which $\hat{Q}$ admits a steady state solution that is rightgoing but not strictly rightgoing. This in turn divides into two principal subcases corresponding to positions (5)-(6) and (7) of Table 2.1.

For the remainder of §4, we will mainly pursue the interpretation of instability as the existence of a *spontaneous rightgoing mode*. Unlike the *reflection coefficient* interpretation, this one corresponds exactly to the GKS stability criterion. It is also relatively easy to make rigorous.

Throughout this discussion our philosophy is that instability need not be studied only abstractly, for it is mainly caused by simple physical mechanisms. By concentrating on these mechanisms we can show that most GKS-unstable difference schemes are susceptible to unstable growth in the ℓ_2 norm (Thms. 4.2.3, 4.2.4), not just in the much less natural GKS norm (Thm. 4.3.1). In the process of isolating this strongly unstable case, we also come to better understand the borderline cases for which the situation regarding stability is less clear.

4.2 ℓ_2-stability; growth theorems

We will consider stability for a general difference model of an initial boundary value problem for a hyperbolic system of equations, as described in §2.5 and §3.6. For much of what follows we could use exactly the formulation of those sections, but to make contact with the GKS stability definition, it is necessary to include in the model an inhomogeneous forcing function $F(x,t)$ and inhomogeneous boundary data $g(t)$.

Consider then the first-order system (cf. (2.5.1))

$$\frac{\partial}{\partial t}u(x,t) = A\frac{\partial}{\partial x}u(x,t) + F(x,t) \qquad (4.2.1)$$

on the quarter-plane $x, t \geq 0$, where $u(x,t)$ and $F(x,t)$ are N-vectors and A is a constant $N \times N$ matrix. Let (4.2.1) be modeled in $x > 0$ by a fixed $s+2$-level difference formula as in (2.5.2), but with the inhomogeneous term added:*

$$Q_{-1}v_j^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma v_j^{n-\sigma} + kF(jh, nk), \qquad j \geq \ell. \qquad (4.2.2)$$

We let Q denote the homogeneous part of this formula (i.e. with $F \equiv 0$), and we assume that Q is Cauchy stable and that it satisfies Assumptions 2.1 (diagonalizability) and 3.2 (n_r, n_ℓ). If (4.2.2) is applied for $j \geq \ell$, then boundary formulas are required to determine values v_j^n for $j = 0, \ldots, \ell-1$, as in (3.6.1). These will be of the form (3.6.1), but with the inhomogeneous term g added:

$$\sum_{j=0}^{j_{max}} \sum_{\sigma=-1}^{\sigma_{max}} S_{j\sigma} v_j^{n-\sigma} = g^n, \qquad (4.2.3)$$

where g^n is a vector of length n_r. For initial conditions, we assume a set of formulas

$$v_j^\sigma = f_j^\sigma, \qquad 0 \leq j < \infty, \quad 0 \leq \sigma \leq s. \qquad (4.2.4)$$

*To get a higher order of accuracy, one might wish to represent F in the model in a more complicated way. This is no problem for the stability theory; see [Go81].

The entities $\{S_{j\sigma}\}$, $\{g^n\}$, and $\{f_j^\sigma\}$ incorporate approximations of all of the boundary or initial data that together with (4.2.1), make up the physical problem to be modeled. $\{S_{j\sigma}\}$ includes in addition any purely numerical boundary conditions. We let the symbol $\tilde{Q}$ denote the complete difference model, (4.2.2)–(4.2.4).

We assume that the following solvability property holds, the natural extension of Ass. 3.3 to inhomogeneous boundary data:

Assumption 4.1. The model $\tilde{Q}$ can be solved boundedly in the sense that if $v^{n-\sigma_{max}}, \ldots, v^n \in \ell_2$ and g^n are given, then v^{n+1} is uniquely determined, and it satisfies a bound

$$\|v^{n+1}\|_2^2 \leq M^2\left(\sum_{\sigma=0}^{\sigma_{max}} \|v^{n-\sigma}\|_2^2 + h|g^n|_2^2\right),$$

where the norms $\|\cdot\|_2$ and $|\cdot|$ are defined as in (3.6.2). //

In setting up the problem we have made three important simplifications. We have left out

(i) variable coefficients $A = A(x,t)$;

(ii) grid-dependent formulas $Q_\sigma = Q_\sigma(k, h(k))$;

(iii) undifferentiated term Hu in (4.2.1).

An important feature of the GKS theory is that it extends to problems with these complications, and although we will discuss only the simplified problem without them, we believe that the same is true for our own arguments based on wave propagation. However, one effect of (i) and (iii) should not be ignored, and that is that they make it possible for solutions to (4.2.1) to grow exponentially with t. Therefore in rewriting the definition of Cauchy stability from §2.2 and §2.5 for initial boundary value problems, we recognize this possibility explicitly, following Defn. 3.1 of [Gu72]:

Defn. Let $\tilde{Q}$ be applied with homogeneous boundary and forcing data, $g \equiv F \equiv 0$. We say that $\tilde{Q}$ is **ℓ_2-stable** if there exist constants $\alpha_0 \geq 0$ and $M > 0$ such that, for all $\alpha > \alpha_0$, the following estimate holds for all $n \geq 0$ and all sufficiently small k:

$$\|e^{-\alpha t}v^n\|_x^2 \leq M \sum_{\sigma=0}^{s} \|f^\sigma\|_x^2 \qquad (t = nk). \qquad (4.2.5)$$

Here $\|\cdot\|_x$ denotes the ℓ_2 norm (3.6.2). //

The definition permits an exponential growth of the solution—at a rate $e^{\alpha_0 t}$, however, that does not increase as the mesh is refined.

We are now in a position to identify mechanisms that can render a difference scheme ℓ_2-unstable. The first important mechanism is Cauchy instability. If the interior formula Q *is not Cauchy stable*, then it cannot satisfy a bound (4.2.5), and easy Fourier arguments show that then $\dot{Q}$ cannot be ℓ_2-stable either. But we have assumed that Q is Cauchy stable.

The second important mechanism was studied by Godunov and Ryabenkii in the early 1960's [Ri67]. Since $\dot{Q}$ does not extend into $x < 0$, a solution of the form $z^n \kappa^j j^\delta \psi$ (2.5.8) with $|\kappa| < 1$ belongs to ℓ_2 for each n. If such a solution exists with $|z| > 1$, then once again we have exponential growth and therefore ℓ_2-instability. One can think of this as spontaneous radiation from the boundary of *strictly rightgoing* energy of type (9) in Table 2.1, that is, of a signal of the kind illustrated in Fig. 2.2a.

For this kind of instability the boundary is definitely involved, and we know that the boundary can couple various wave components ψ_ν (§2.5). Therefore in general we must look not just for one solution $z^n \kappa^j j^\delta \psi$, but for linear combinations of such modes. We define:

Defn. Let $z \in \mathbb{C}$ satisfy $|z| \geq 1$, and suppose $\dot{Q}$ with $F \equiv g \equiv 0$ admits as a solution a linear combination of rightgoing modes

$$v_j^n = z^n \phi_j = z^n \sum_{\nu=1}^{q} a_\nu \kappa_\nu^j j^{\delta_\nu} \psi_\nu, \qquad a_\nu \neq 0 \tag{4.2.6}$$

as defined in (2.5.10), *where for each* ν, $|\kappa_\nu| < 1$. Then ϕ is an **eigensolution** of $\dot{Q}$ with **eigenvalue** z. (Eigensolutions with $|z| < 1$ can also readily be defined, but these are not relevant to stability.) //

In other words, an eigensolution is a linear combination of signals from position (7) of Table 2.1 in the case $|z| = 1$, or from position (9) in the case $|z| > 1$, that satisfies both the homogeneous interior formula (4.2.2) and the homogeneous boundary conditions (4.2.3). (We will abuse terminology by referring to both ϕ and $z^n\phi$ as eigensolutions, as convenient.) We define further:

Defn. A **strictly rightgoing eigensolution** is an eigensolution consisting entirely of strictly rightgoing signals. Equivalently, it is an eigensolution with $|z| > 1$ (position (9) of Table 2.1). //

The Godunov-Ryabenkii theorem now states:

Theorem 4.2.1 (Godunov-Ryabenkii theorem) *[Ri67]. A necessary condition for ℓ_2-stability of $\dot{Q}$ is that there exist no strictly rightgoing eigensolution.*

Proof. Suppose there exists a strictly rightgoing eigensolution ϕ. If $v_j^\sigma = z^\sigma \phi_j$ is taken as initial data (4.2.4) for $0 \leq \sigma \leq s$, the solution as n increases will be $v_j^n = z^n \phi_j$ for all n. Since $t = nk$, this means that v will grow like $|z|^{t/k}$. This growth is unbounded for any t as $k \to 0$, which contradicts (4.2.5). ∎

This theorem has a direct restatement in terms of the reflection matrix $D^{[r]}$ of §3.6:

Theorem 4.2.2 (Godunov-Ryabenkii theorem, determinant condition). *A necessary condition for ℓ_2-stability of $\dot{Q}$ is that for all z with $|z| > 1$, the matrix $D^{[r]}$ of (3.6.4) is nonsingular, i.e.*

$$\det D^{[r]}(z) \neq 0 \qquad \textit{if } |z| > 1.$$

Proof. If $D^{[r]}(z)$ is singular for some z with $|z| > 1$, let $a^{[r]}$ be a corresponding homogeneous right eigenvector. Then the function

$$\phi_j = \sum_{\nu=1}^{n_r} a_\nu^{[r]} \kappa_\nu^j j^{\delta_\nu} \psi_\nu, \tag{4.2.7}$$

as in (2.5.10), is an unstable strictly rightgoing eigensolution. ∎

The limitation of the Godunov-Ryabenkii condition is that although it is necessary for stability, it is far from sufficient, both in theory and in practice. What it fails to take into account is a third instability mechanism, namely the existence of strictly rightgoing *wavelike* solutions (i.e. with $|z| = |\kappa| = 1$). For this we make use of the concept of a *generalized eigensolution*, which was introduced by Kreiss but is defined here from our wave propagation point of view:

Defn. Let $z \in \mathbb{C}$ satisfy $|z| = 1$, and suppose $\dot{Q}$ with $F \equiv g \equiv 0$ admits as a solution a linear combination of rightgoing modes

$$v_j^n = z^n \phi_j = z^n \sum_{\nu=1}^{q} a_\nu \kappa_\nu^j j^{\delta_\nu} \psi_\nu, \qquad a_\nu \neq 0 \tag{4.2.8}$$

as defined in (2.5.10), where for at least one ν, $|\kappa_\nu| = 1$. Then ϕ is a **generalised eigensolution** of $\dot{Q}$ with **generalised eigenvalue** z. //

In analogy with the earlier definition we now state:

Defn. A **strictly rightgoing generalized eigensolution** is a generalized eigensolution consisting entirely of strictly rightgoing signals. Equivalently, it is any generalized eigensolution with $|\kappa_i| = 1$ and $C_i > 0$ for all i. //

This definition leads to the following theorem, which is new. Let S denote the multilevel solution operator for the homogeneous model $\tilde{Q}$ with $g \equiv F \equiv 0$:

$$S : \{v^n, \ldots, v^{n-s}\} \mapsto \{v^{n+1}, \ldots, v^{n-s+1}\}. \tag{4.2.9}$$

Let these $s+1$-level vectors be normed by

$$\|\{v^n, \ldots, v^{n-s}\}\|_2^2 = \sum_{\sigma=0}^{s} \|v^{n-\sigma}\|_2^2, \tag{4.2.10}$$

with the norm on the right defined by (3.6.2), and let $\|S\|_2$ be the induced operator norm.

Theorem 4.2.3. *A necessary condition for ℓ_2-stability of $\tilde{Q}$ is that there exist no strictly rightgoing generalized eigensolution. If there does exist a strictly rightgoing generalized eigensolution, then*

$$\|S^n\|_2 \geq \text{const.} \sqrt{n} \tag{4.2.11}$$

for infinitely many integers $n \geq 0$.

Proof. See Appendix B. ∎

The proof of this theorem has been deferred to an appendix for clarity here. However, the explanation of the result is exactly what was discussed in §4.1. If the initial data consist of a narrow signal at the boundary of the form of the generalized eigensolution, then as time elapses it will move steadily rightward, as suggested in Fig. 4.3.

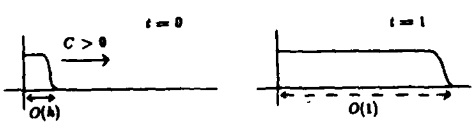


FIG. 4.3

As a result the solution grows in ℓ_2 as fast as $\sqrt{n}$. Precisely this argument can be made rigorous, but for technical simplicity the proof in App. B proceeds somewhat differently.

Whether (4.2.11) captures the rate of growth observed in practice for an unstable difference model appears to depend on *reflection coefficients*. In Demo. 4.2 we saw that if an infinite reflection coefficient is present, then amplitude growth may be observed that is proportional to n, not $\sqrt{n}$. Therefore we propose:

Conjecture. *The bound (4.2.11) is sharp in the sense that there are some ℓ_2-unstable models $\tilde{Q}$ admitting strictly rightgoing generalized eigensolutions for which*

$$\|S^n\|_2 \leq \text{const.} \sqrt{n} \quad \forall n > 0. \tag{4.2.12}$$

However, suppose that $\tilde{Q}$ has a strictly rightgoing generalized eigensolution for which the reflection matrix $[D^{[r]}(z)]^{-1} D^{[\ell]}(z)$ is infinite. Then (4.2.11) can be strengthened to

$$\|S^n\|_2 \geq \text{const.}\ n \quad \forall n > 0. \tag{4.2.13}$$

In addition to stability with respect to initial data f, it makes sense to consider stability with respect to forcing data F or boundary data g. Our proof of Thm. 4.2.3 can in fact be used to show that a bound analogous to (4.2.11) holds for problems driven by F. Probably the natural analogs of (4.2.12) and (4.2.13) hold also. For boundary data, however, the situation is different - we get growth proportional to n regardless of the reflection coefficients. Let $\tilde{Q}$ be applied with $f \equiv F \equiv 0$ but with $g \not\equiv 0$. Let $S_{bc}^{(n)}$ denote the operator

$$S_{bc}^{(n)} : g \mapsto v^n \tag{4.2.14}$$

with norm induced by ℓ_2 norms for g and v^n with respect to t and x, respectively.

Theorem 4.2.4. *A necessary condition for stability of $\tilde{Q}$ with respect to boundary data is that there exist no strictly rightgoing eigensolution or generalized eigensolution. If there does exist such a solution, then*

$$\|S_{bc}^{(n)}\|_2 \geq \text{const.}\ n \quad \forall n > 0. \tag{4.2.15}$$

Proof. See Appendix B. ∎

There is little doubt that as with (4.2.13), this bound is sharp for the class of strictly rightgoing generalized eigensolutions as a whole, although faster growth can be obtained in particular cases.

• • •

It is natural to ask whether the growth rates (4.2.11), (4.2.13), (4.2.15) are severe enough to cause trouble in practice. For the latter two cases (linear growth) the answer is clearly yes. Whatever the problem being solved, rightgoing radiation at the boundary will tend to appear in these cases, causing the computation to give unreasonable answers. As a minimum it will result in failure to converge as the grid is refined. The numerical examples of the next chapter will illustrate these claims (see especially Figs. 5.2b, 5.4.3 b). For the question of stability with respect to initial data in the finite reflection coefficient case (4.2.11), however, the situation is more delicate. We will give evidence in Chapter 5 that the instability here is quite weak in practice.

There is another important justification for considering the kind of growth we have described unstable, which is often mentioned by Kreiss. That is, if a ***second boundary*** is introduced in the problem being modeled, say at $x = 1$, its effect may be to convert an algebraic growth rate to exponential. If one hopes for a stability theory that permits one to investigate the stability of each boundary individually, it follows that a model with a strictly rightgoing generalized eigensolution will have to be considered unstable. However, we will discuss problems involving two boundaries at length in §5 and §6.5, and conclude that the exponential growth occurs only if the unstable boundary has an infinite reflection coefficient.

4.3 GKS-stability

Theorems 4.2.1 and 4.2.3 give necessary but not sufficient conditions for stability. As has been stated, we believe that in practice these conditions are more or less sufficient also, at least for stability with respect to initial data, and we will give various examples in support of this view in §5. However, no estimate on the growth of v is available to make this opinion precise. In fact in at least one (quite contrived) situation, these conditions are demonstrably too weak to ensure ℓ_2-stability. This

is the case in which $\tilde{Q}$ admits an eigensolution with $|z| = 1$, but in which z is a defective eigenvalue of $P_\kappa(z)$ (§2.5) and it is also defective with respect to the boundary conditions. In this event Thm. 2.1.1 implies that one must expect algebraic growth with n.

A striking achievement of the GKS theory is that it obtains a ***necessary and sufficient*** condition for stability. This is accomplished by extending the stability conditions of Thms. 4.2.1 and 4.2.3 to include non-strictly rightgoing solutions, and by strengthening the definition of stability. Here is the new definition, which appears as Defn. 3.3 in [Gu72]:

Defn. Let $\tilde{Q}$ be applied with homogeneous initial data $f \equiv 0$. We say that $\tilde{Q}$ is **GKS-stable** if there exist constants $\alpha_0 \geq 0$ and $M > 0$ such that, for all $\alpha > \alpha_0$, the following estimate holds for all sufficiently small k:

$$\left(\frac{\alpha - \alpha_0}{1 + \alpha k}\right)^2 \|e^{-\alpha t} v\|_{x,t}^2 + \left(\frac{\alpha - \alpha_0}{1 + \alpha k}\right) \sum_{j=0}^{\ell - 1} \|e^{-\alpha t} v_j\|_t^2$$
$$\leq M\left(\|e^{-\alpha(t+k)} F\|_{x,t}^2 + \left(\frac{\alpha - \alpha_0}{1 + \alpha k}\right) \|e^{-\alpha(t+k)} g\|_t^2\right). \tag{4.3.1}$$

Here $t = nk$ and $\|\cdot\|_{x,t}$ and $\|\cdot\|_t$ denote the ℓ_2 norms defined by

$$\|\phi\|_{x,t}^2 = hk \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} |\phi_j^n|^2, \qquad \|\phi\|_t^2 = k \sum_{n=0}^{\infty} |\phi_j^n|^2. \tag{4.3.2}$$

This definition is quite forbidding, and some remarks on it are in order:

(1) Unlike (4.2.5), the bound (4.3.1) involves F (and g) rather than f. This is an unfortunate technical limitation that is made necessary by the proofs of the GKS theorems, which are based on a Fourier transform in t. If (4.3.1) involved f but not F, then one would be able to extend it to a bound involving F also by means of the discrete analog of Duhamel's principle. The connection in the other direction is however not so easy; the obvious approach requires the introduction of a factor $1/k$ in the right hand side of (4.3.1). For the problem of well posedness of partial differential equations (as opposed to difference models), by contrast, a complete connection between f and F is known to hold [Ra72].

(2) The Fourier transform arguments are also responsible for the appearance of decay factors $e^{-\alpha t}$ on the right as well as the left, and for the normalizing fractions $(\frac{\alpha - \alpha_0}{1 + \alpha k})$. Like (4.2.5), (4.3.1) permits exponential growth at the rate $e^{\alpha_0 t}$.

(3) The boundary term $\sum \|e^{-\sigma t} v_j\|_t^2$ gives special weight to the behavior of the solution near $x = 0$. This is an important point that we will discuss below and in §5.

(4) A valuable property of this definition is that one can show that the set of GKS-stable difference schemes is *open* in the following sense: if $\tilde{Q}$ is stable, then a perturbed scheme $\hat{Q}$ is stable also, provided $\|\hat{Q} - \tilde{Q}\| = O(k)$ as $k \to 0$ (Thm. 4.3 of [Gu72]). It is this robustness that makes the GKS theory extend readily to problems with the complications (i)-(iii) listed in §4.2, and also to problems with two boundaries.

Because of (3), GKS-stability is a substantially more stringent requirement than ℓ_2-stability. However, it is not known whether GKS-stability actually implies ℓ_2-stability. Kreiss et al. conjecture in §3 of [Gu72] that it does.

The main GKS theorem is like Thms. 4.2.1 and 4.2.3, except that the hypothesis of a strictly rightgoing mode is removed and an additional dissipativity restriction is added:

Theorem 4.3.1 (GKS stability theorem). *Assume that Q is either x-dissipative or strictly nondissipative.* A necessary and sufficient condition for GKS-stability of $\tilde{Q}$ is that there exist no rightgoing eigensolution or generalized eigensolution (i.e., no eigensolution or generalized eigensolution with $|z| \geq 1$).*

Proof. This theorem is equivalent to Lemma 10.3 and the sentence following in [Gu72]. ∎

The proof given in [Gu72] is a lengthy one, and to prove that the eigensolution condition is sufficient for stability, we know of no alternatives. But as in Thm. 4.2.3, the necessity can be established by arguments of dispersive wave propagation. We have stated that the essential feature of the GKS stability definition is the integral along $x = 0$ that it includes. The following argument will work for any stability definition involving such a boundary integral.

Sketch of proof of necessity in Thm. 4.3.1. As in the proof of Thm. 4.2.3, suppose $\tilde{Q}$ admits a rightgoing solution (4.2.6). Once again, we want to construct an initial signal consisting of this solution for x near 0, cutting off smoothly to $v \equiv 0$ near $x = \epsilon$, as in Fig. 4.3a. (Since the GKS-stability definition involves F rather than f, this signal must be introduced through F rather than f.) Now as t increases, each wave front in (4.2.6) remains stationary or moves right. In either event the initial signal sits essentially unchanging near the origin. Because of the boundary term $\sum \|e^{-\sigma t} v_j\|_t^2$ on the left of (4.3.1), this stationary behavior can be seen to be GKS-unstable.

*This is Assumption 5.4 of [Gu72]. It appears to be unknown to what extent this restriction is necessary for the theorem to go through. We conjecture that for diagonalisable difference models, at least, it is unnecessary. Osher's results of [Os69b] show this is true for at least some problems.

The GKS theorem has a simple restatement in terms of a determinant condition (cf. Thm. 4.2.2):

Theorem 4.3.2 (GKS stability theorem, determinant condition). *A necessary and sufficient condition for GKS-stability of $\tilde{Q}$ is that for all z with $|z| \geq 1$, the reflection matrix $D^{[r]}$ of (3.6.4) is nonsingular, i.e.*

$$\det D^{[r]}(z) \neq 0 \qquad \text{if } |z| \geq 1.$$

Proof. The determinant condition is equivalent to the condition of Thm. 4.3.1, by the same argument as in the proof of Thm. 4.2.2. ∎

• • •

To summarize §4.2 and §4.3, we have shown that unstable difference models of initial boundary value problems can be recognized by the unstable steady-state solutions they admit. If $\tilde{Q}$ admits a strictly rightgoing solution, it is unstable in ℓ_2 with a growth rate of at least $\sqrt{n}$, and probably n when an infinite reflection coefficient is present. If it admits a rightgoing solution with no strictly rightgoing components, it is still unstable according to the GKS definition. Since the definition of "rightgoing" for wavelike modes depends on the group velocity, these results demonstrate that group velocity has a fundamental role in determining stability.

We have not mentioned stability for problems with interfaces, except to fold them into initial boundary value problems. However, the results above unfold easily, and we find: *an interface model is unstable if it admits a steady-state solution that is* **outgoing** *from the point of view of the interface* (leftgoing on the left, rightgoing on the right).

We have also not mentioned the "perturbation test for generalized eigensolutions," which is described in various accounts of the GKS results, but which many practitioners find mysterious. This is nothing more than the perturbation test for distinguishing positive and negative group velocities that was described in Thm. 2.3.2.

4.4 Stability for dissipative schemes

All of the statements of the past three sections apply to dissipative formulas, for nowhere have we assumed nondissipativity. In particular, recall that Thm. 2.3.1 guarantees that the group velocity makes sense for any mode with $|z| = |\kappa| = 1$, even if it is admitted by a dissipative formula. However, it is worth discussing dissipative models explicitly, both because the stability criteria can be simplified in this case, and because dissipative models are a natural point of confusion regarding the validity and scope of the group velocity approach to stability.

Suppose first that the interior formula Q is ***totally dissipative***, which means that $|\kappa| = |z| = 1$ is possible only for $\kappa = z = 1$ (§1.2). From Table 2.1, it is evident that this restricts the set of rightgoing solutions admitted by Q, apart from $z = \kappa = 1$, to the possibilities $|z| = 1 > |\kappa|$ and $|z| > 1 > |\kappa|$. From the definitions of eigensolutions and generalized eigensolutions in §4.2, it follows that the GKS theorem (Thm. 4.3.1) takes the following special form:

Theorem 4.4.1 *(GKS theorem for totally dissipative schemes). Let Q be totally dissipative. A necessary and sufficient condition for GKS-stability of $\tilde{Q}$ is that the following conditions hold:*

(i) There are no (rightgoing) eigensolutions with $|z| \geq 1$;

(ii) There are no (rightgoing) generalized eigensolutions that involve the wave mode $\kappa = z = 1$. ∎

Similar special formulations could be given for the theorems of §4.2.

The advantage of this statement over Thm. 4.3.1 is that it enables one to limit the search for unstable wavelike modes to the single point $\kappa = z = 1$. This point is special, of course, in that it corresponds to the partial differential equation being modeled whenever $\tilde{Q}$ is a consistent approximation. Therefore one is tempted to rewrite condition (ii) above as the condition that $\tilde{Q}$ is consistent with a well-posed initial boundary value problem. However, this is not strong enough, because, for example, of the possibility of an unstable rightgoing solution consisting of some energy in the mode $\kappa = z = 1$ *plus* additional energy in a component with $z = 1$, $|\kappa| < 1$.

As mentioned in §0.2, much of the early work on stability for models of initial boundary value problems was confined to the case in which Q is a two-level x-dissipative formula, hence by Thm. 2.2.3, totally dissipative. Therefore the point $\kappa = 1$ takes on a special significance in these papers. The results derived in them are not necessary and sufficient conditions, but they have the considerable advantage of being stated for the ℓ_2 norm. In particular, the main theorems of [Kr66], [Os69a], and [Kr68] state approximately that for two level dissipative models, conditions *(i)* and *(ii)* above are sufficient for ℓ_2-stability.

Now suppose that Q is x-dissipative but not necessarily t-dissipative (§1.2). This possibility comes up only for multilevel schemes, such as LFd (Leap Frog with dissipation, §1.1). Then Thm. 4.4.1 holds if *(ii)* is replaced by the condition that there is no rightgoing generalized eigensolution involving a wave mode with $|z| = 1$, $\kappa = 1$. This class of problems has not received separate treatment in the literature.

These results naturally lead to an important question: does one need dissipativity in order to be able to derive theorems involving ℓ_2-stability instead of just GKS-stability? The results in §4.2 have shown that for necessary conditions one does not, but this leaves open the matter of sufficient conditions. Our belief is that although ℓ_2 results may be harder to derive in the nondissipative case, there is no reason why they should be unobtainable. In fact, it has already been mentioned in §4.3 that it is likely that GKS-stability implies ℓ_2 stability, in which case Thm. 4.3.1 provides one such sufficient condition—although, as will be shown in §5.4, it is not sharp.

Finally, what if Q is neither x-dissipative nor totally dissipative? For example, Q might be a t-dissipative scheme such as BE or LxF (App. A), admitting a finite collection of wavelike modes rather than a continuum of them. It is certainly likely that Thm. 4.4.1 is valid for these problems, if condition *(ii)* is extended in the obvious way to cover all points where $|\kappa| = |z| = 1$ is possible under Q. In fact, Osher's results of [Os69b] show stability for BE and LxF with certain kinds of boundary conditions (see especially XXIII of [Os69b]). However, as mentioned in the last section, the proofs of [Gu72] do not cover this case.

• • •

In addition to these theoretical remarks, there is a practical point to be mentioned: totally dissipative models are much less often unstable than nondissipative ones. In practice, despite the qualification above, one rarely encounters instabilities of type *(ii)* in Thm. 4.4.1, so that this leaves the possibility of eigensolutions (i). For simple problems in one dimension, these almost never appear unless one is looking for them, so that as a rule, one can usually make an unstable model Q stable by adding some dissipation. However, as the complexity of the problem goes up, and especially if more than one space dimension is involved, the rule "totally dissipative implies stable" becomes less and less reliable.

An example will suffice to show that even for very simple problems, one can devise unstable totally dissipative models. Let $u_t = u_x$ be modeled by LW with $\lambda = 1/3$ *for* $j \geq 1$, together with the boundary formula

$$v_0^{n+1} = v_0^n + \lambda(v_2^n - v_1^n). \tag{4.4.1}$$

One readily verifies that this scheme admits a strictly rightgoing eigensolution of Godunov-Ryabenkii type: $z = 31/27$, $\kappa = -1/3$. Numerical experiments confirm that any solution attempted with this scheme is rapidly obliterated by noise growing at the rate $(31/27)^n$. However, note how contrived the condition (4.4.1) is — it would never be proposed in practice.

Sections 6.2 and 6.3 investigate the connection between dissipativity and stability further for some boundary and interface problems.

4.5 Some general classes of unstable difference models

In practice, as we have mentioned, a large proportion of instabilities that appear in difference models of initial boundary value problems are not eigensolutions but generalized eigensolutions. *Within the range of generalized eigensolutions,* it turns out further that in practice, a large proportion of instabilities involve simple sawtoothed waves with $z = -1$ and/or $\kappa = -1$. (Analogously, when a difference model for an initial value problem is unstable, it is usually an unstable sawtoothed mode that dominates.) As we saw in §1, sawtoothed modes are by no means the only waves that travel in the physically wrong direction. The reason for their predominance in practice is that other waves which do so, for which κ and z have values on the unit circle other than ± 1, do not as often satisfy the numerical boundary conditions.

It was with the significance of sawtoothed parasites in mind that we defined the *concepts of x- and t-reversing* difference formulas in §1.5. We can now apply these definitions to delineate some general classes of unstable difference models. All of the theorems in this section are new, but they are straightforward generalizations of well known examples. One purpose in collecting them together is to demonstrate that once the stability question for initial boundary value problems is given a physical meaning, it becomes natural to consider difference schemes in groups rather than one by one.

1: space extrapolation with t-reversing formulas

Let $u_t = u_x$ be modeled by a difference formula Q for $j \geq \ell$ coupled with $(q_j - 1)$*st-order space extrapolation* boundary conditions (cf. (3.2.29))

$$S: \qquad [(K-1)^{q_j} v^{n+1}]_j = 0 \qquad (0 \leq j \leq \ell - 1) \tag{4.5.1}$$

for the boundary points $j < \ell$, with $q_j \geq 1$ for each j. For the case of $Q = LF$ and $\ell = q_0 = 1$, we showed in §4.1 that this scheme *admits the unstable strictly* rightgoing mode $(\kappa, z) = (1, -1)$, and the same result has appeared in [Gu72, §6] and in various other places.

Here is a natural generalization:

Theorem 4.5.1. *Any consistent t-reversing difference formula Q for (1.1.1) is ℓ_2- and GKS-unstable in combination with the boundary condition S.*

Proof. Assume first $a > 0$. The sawtoothed wave $v_j^n = (-1)^n$ satisfies S for any set $\{q_j\}$, and if Q is t-reversing, it also satisfies Q and has $C > 0$, since by consistency $v_j^n \equiv 1$ must satisfy Q with $C = -a < 0$. By Thms. 4.2.3 and 4.3.1, the model is therefore ℓ_2- and GKS-unstable. For $a < 0$, on the other hand, $v_j^n \equiv 1$ is itself an unstable rightgoing mode. (In this case the model is not consistent with any well-posed differential equation.) ∎

This is an example in which the reflection coefficient for the unstable mode is infinite, as was pointed out in §4.1, so that growth like $\|S^n\| \geq$ const. n can be expected. Thm. 4.5.1 applies even for schemes that are x- but not t-dissipative, such as LFd or various analogous schemes consisting of LF4 with spatial dissipation added. The instability of S with LFd has been pointed out by Goldberg and Tadmor in Example 4.1 of [Go81]. One can also readily extend Thm. 4.5.1 to t-reversing formulas in combination with arbitrary extrapolation boundary conditions, provided that they are at least zeroth order accurate and confined to a single time level.

2: "one-sided leap frog" with t-reversing formulas

Similarly, it has been noted in various papers that if (1.1.1) is modeled by LF for $j \geq 1$ together with the boundary condition

$$v_0^{n+1} = v_0^{n-1} + 2\lambda a(v_1^n - v_0^n),$$

then the result is GKS-unstable. As a generalization, consider any set of boundary conditions

$$v_j^{n+1} = v_j^{n-1} + 2ka D_j v_j^n \qquad 0 \leq j \leq \ell - 1, \tag{4.5.2}$$

where each D_j is a spatial difference operator consistent with $\partial/\partial x$ that involves at most j points to the left of center. We obtain just as above

Theorem 4.5.2. *Any consistent t-reversing difference formula Q for (1.1.1) is ℓ_2- and GKS-unstable in combination with the boundary condition (4.5.2).*

Proof. Same as for Thm. 4.5.1. ∎

3: sign-changing coefficients; nonlinear instability

Consider the coefficient-change problem (3.2.1). As in Example 3.1, suppose we model this on a grid (jh, nk) by consistent difference formulas Q_- for $j \leq -1/2$ and Q_+ for $j \geq 1/2$, respectively. According to (3.2.5) or (3.2.6), the reflection and transmission coefficients will become infinite in this problem if there exists a steady-state solution in which $\kappa_* = \kappa_t$, that is, a uniform wave that is leftgoing on the left and rightgoing on the right. If $\operatorname{sgn} a_- = \operatorname{sgn} a_+$, then most models do not admit such solutions, and they are stable. But stability vanishes if $\operatorname{sgn} a_- \neq \operatorname{sgn} a_+$.

Theorem 4.5.3. *Let (3.2.1) be modeled by consistent formulas Q_- and Q_+ as indicated above. If $a_- > 0 > a_+$, the model is ℓ_2- and GKS-unstable. If $a_- < 0 < a_+$ and Q_- and Q_+ are both x-reversing or both t reversing, the model is again ℓ_2- and GKS unstable.*

Proof. In the first case, the constant function $v_j^n \equiv 1$ is an outgoing wave that satisfies all of the difference formulas, so the model is unstable by Thms. 4.2.3 and 4.3.1. In the second case, the same goes for a space or time sawtooth $(-1)^j$ or $(-1)^n$. ∎

This elementary example is related to certain known examples of *nonlinear instability*. If the Burgers equation

$$u_t = uu_x$$

is modeled by the leap frog scheme

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

then exponentially growing instabilities arise that are marked by oscillations of the form [Fo73,Kr73]

$$v_j^n \approx 0, \quad v_{j+1}^n < 0, \quad v_{j+2}^n > 0, \quad v_{j+3}^n \approx 0.$$

Though it is easy enough to examine this problem directly, it also has a rough interpretation along wave propagation lines. LF is an x reversing formula, and the instability observed looks approximately like the outgoing spatial sawtooth $(-1)^j$ of Thm. 4.5.3 from the point of view of the sign-change interface at $x_{j+3/2}$. The linear growth of this outgoing wave would be converted to exponential by reflection at points x_j and x_{j+3} even if the coefficients v_j did not change from one time step to the next; the fact that they do makes the growth still more rapid.

For an interesting study of a nonlinear instability with a more subtle explantion related to wave propagation, see the paper [Br81] (especially §4) by Briggs, et al.

4: "coarse mesh" mesh refinement

Consider the "coarse mesh approximation" mesh refinement scheme of Example 3.4, in which a three-point linear multistep formula is applied with space step h_- for $x < 0$ and $h_+ = mh_-$ for $x > 0$, with the formula (3.2.22) imposed at the interface. According to (3.2.23), this scheme possesses an infinite reflection coefficient if there exists a frequency ω for which κ_t (transmitted) $= \kappa_r^m$ (reflected). When m is even, this situation can easily occur. The following theorem generalizes the setup somewhat:

Theorem 4.5.4. *Let (1.1.1) be modeled by a consistent x reversing 3 point formula Q_- on $x_j = jh$ for $j \leq -1$ coupled with any consistent formula Q_+ on $x_j = jmh$ for $j \geq 0$, with left hand values for the latter near the interface taken where needed from points imh with $i \leq -1$. If $a < 0$ and m is even, the model is ℓ_2- and GKS-unstable. If $a > 0$ and m is even and both Q_- and Q_+ are t reversing, the model is again ℓ_2- and GKS unstable.* ∎

Proof. In the case $a < 0$, consider a wave

$$v_j^n = \begin{cases} (-1)^j & (j \leq 0), \\ 1 & (j \geq 0). \end{cases} \tag{4.5.3}$$

On $x \geq 0$, this wave is constant and has $C = -a > 0$. On $x \leq 0$, it is sawtoothed and has $C < 0$ since Q_- is x-reversing. Thus (4.5.3) is outgoing on both sides of the interface. Moreover *if m is even*, it obviously satisfies the boundary formulas, so we have instability. In the case $a > 0$, multiply (4.5.3) by $(-1)^n$. ∎

For LF, CN, and many other formulas, the sawtooths we have considered turn out to be the only instabilities that arise, so this mesh refinement scheme is stable when m is odd (Joseph Oliger, private communication).

4.6 Unstable difference schemes in several space dimensions

In the study of well-posedness of hyperbolic partial differential equations on a region with a boundary, problems in one space dimension are easy to treat (by the method of characteristics), but in two or more space dimensions the situation becomes complicated. By a multidimensional problem, we have in mind an equation defined on the d-dimensional half space $x_1 \geq 0$, $x_j \in \mathbb{R}$ for $2 \leq j \leq d$. The main theory available for this was derived in an important paper by Kreiss in 1970 [Kr70], by techniques that formed the basis of the GKS theory for difference models published two years later [Gu72]. Like the stability criteria that we have discussed in §§4.2–4.3, Kreiss's well-posedness criterion is a determinant condition that requires, roughly, that the problem admit no spontaneous rightgoing signals at the boundary. The difference is that for the differential equation, the question of whether a signal is rightgoing depends on multidimensional geometric effects, but does not involve a nontrivial group velocity, since the system is nondispersive. Similarly, it is well known that hyperbolic equations in more than one space dimension are generally ill-posed in L_p norms for $p \neq 2$, as we have seen for finite-difference models in one space dimension (§1.4), but this is due to geometric focusing rather than dispersion.

For finite difference models in two or more dimensions, focusing and dispersion effects are combined. The corresponding stability theory has been late in appearing. Some results follow from the one-dimensional theory by a Fourier transform in the variables $x_2, \dots, x_d$, but these were never developed by Kreiss, et al. See also the paper [Os69c] by Osher. More recent results in this area are due to Coughran [Co80] and especially Michelson [Mi81]. Both of these authors consider only difference schemes that satisfy a dissipativity condition — in the former case, one that is related to our definition of t-dissipativity (§2.2).

Our purpose in this section is to point out that the wave propagation arguments we have developed for one space dimension provide immediate necessary conditions for stability of both dissipative and nondissipative difference models in several dimensions, too.

• • •

We will confine the discussion to a simple class of examples. Abarbanel and Gottlieb [Ab79] and Abarbanel and Murman [Ab81] have studied the stability of various difference schemes for the following problem in two space dimensions:

$$u_t = u_x + u_y \qquad x, t \geq 0, \quad y \in (-\infty, \infty). \tag{4.6.1}$$

The solutions to this equation consist of functions

$$u(x, y, t) = u(x + t, y + t, 0).$$

That is, information propagates with a vector velocity $(-1, -1)$. Since the flow is outward across the boundary $x = 0$, no boundary conditions should be given there.

For a multidimensional problem like this, we saw in §1.6 that ξ becomes a wave number vector $\boldsymbol{\xi}$, and the group speed C generalizes to a vector group velocity given by the gradient

$$C = \nabla_{\xi}\, \omega. \tag{4.6.2}$$

By the same arguments as in Thms. 4.2.3 and 4.3.1, one can readily obtain the following stability result: *if a finite difference model of (4.6.1) admits a solution consisting of waves with group velocity C pointing into $x \geq 0$ (i.e. with $C_x \geq 0$), it is GKS-unstable. If each wave has $C_x > 0$, then it is also ℓ_2-unstable, with a growth rate at least proportional to $\sqrt{n}$.* We will not go to the trouble here of developing the stability definitions in this theorem, or of writing down a proof, because there are no ideas involved that were not present in one dimension.

As an example, suppose (4.6.1) is modeled by the leap frog formula

$$v_{ij}^{n+1} - v_{ij}^{n-1} = \lambda(v_{i+1,j}^n - v_{i-1,j}^n) + \lambda(v_{i,j+1}^n - v_{i,j-1}^n). \tag{4.6.3}$$

The dispersion relation for this scheme is

$$\sin \omega k = -\lambda \sin \xi h - \lambda \sin \eta h,$$

where $\boldsymbol{\xi} = (\xi, \eta)$, and from (4.6.2) there follow the group velocity components

$$C_x = -\frac{\cos \xi h}{\cos \omega k}, \qquad C_y = -\frac{\cos \eta h}{\cos \omega k}.$$

As usual, these reduce to the ideal value $C = (-1, -1)$ for $\xi h, \omega k \approx 0$. If we look at parasites, on the other hand, we see that a sawtooth form in x or y negates C_x or C_y, respectively, and a sawtooth in t negates both. Table 4.1 summarizes the situation:

	$\xi h, \eta h, \omega k$	C
(a)	$(0,0,0), (\pi,\pi,\pi)$	$(-1,-1)$
(b)	$(\pi,0,0), (0,\pi,\pi)$	$(+1,-1)$
(c)	$(0,\pi,0), (\pi,0,\pi)$	$(-1,+1)$
(d)	$(\pi,\pi,0), (0,0,\pi)$	$(+1,+1)$

TABLE 4.1

Thus sawtoothed parasites can travel in any of the directions at 45° to the grid. If any parasite of form (b) or (d) is permitted by the boundary conditions, the difference model is unstable.

Abarbanel et al. consider various boundary formulas. **Four of these are** ***space extrapolation*** and ***skewed space extrapolation*** (cf. (3.2.29)),

S: $\quad (K_x - 1)^q v_0^{n+1} = 0,$

SS: $\quad (K_x K_y - 1)^q v_0^{n+1} = 0,$

and ***space/time extrapolation*** and ***skewed space/time extrapolation*** (cf. (3.2.32)),

ST: $\quad (K_x Z^{-1} - 1)^q v_0^n = 0,$

SST: $\quad (K_x K_y Z^{-1} - 1)^q v_0^n = 0$

Here K_x, K_y, and Z denote the shift operators in x, y, and t. By counting sign changes, one can see which boundary formulas permit which sawtooths. The results are listed in Table 4.2:

	stable sawtooths	*unstable sawtooths*
S	$(0,0,0), (0,\pi,0)$	$(0,0,\pi), (0,\pi,\pi)$
SS	$(0,0,0), (\pi,\pi,\pi)$	$(0,0,\pi), (\pi,\pi,0)$
ST	$(0,0,0), (0,\pi,0), (\pi,0,\pi), (\pi,\pi,\pi)$	
SST	$(0,0,0), (\pi,0,\pi)$	$(0,\pi,\pi), (\pi,\pi,0)$

TABLE 4.2

Thus *S, SS, and SST are all unstable with LF.* It turns out that ST, which the table shows has no sawtooth instabilities, admits no other rightgoing solutions either.

Other difference formulas typically permit fewer sawtooths, hence are stable with more boundary conditions. Let us generalize to d space dimensions. If $\boldsymbol{\kappa}$ and $\boldsymbol{j}$ are d-vectors, $\boldsymbol{\kappa}^{\boldsymbol{j}}$ will denote $\kappa_1^{j_1} \cdots \kappa_d^{j_d}$.

Defn. Let Q be a scalar difference formula in d space dimensions. Suppose that whenever Q admits a solution $v_j^n = \boldsymbol{\kappa}^j z^n$ with $|z| = |\kappa_i| = 1$ for each i, $\kappa_l \neq 1$ for some l, and group velocity $C \in \mathbb{R}^d$, then it also admits the solution $v_j^n = (-1)^{j_l} \boldsymbol{\kappa}^j z^n$, and this wave has group velocity $C' \in \mathbb{R}^d$ satisfying $C'_i = C_i$ for $i \neq l$ and $C_l C'_l \leq 0$, with $C'_l \neq 0$ if $C_l \neq 0$. Then Q is x_l**-reversing**. Suppose that whenever Q admits a solution $v_j^n = \boldsymbol{\kappa}^j$ with $|\kappa_i| = 1$ for all i and group velocity $C \in \mathbb{R}^d$, then it also admits the solution $v_j^n = \boldsymbol{\kappa}^j(-1)^n$, with group velocity $C' \in \mathbb{R}^d$ satisfying $C_i C'_i \leq 0$ for $1 \leq i \leq d$, with $C'_i \neq 0$ if $C_i \neq 0$. Then Q is t**-reversing.** //

Now let Q be a consistent difference model of

$$u_t = \sum_{j=1}^{d} u_{x_j}$$

on $t, x_1 \geq 0$, $x_j \in (-\infty, \infty)$ for $2 \leq j \leq d$, and let the boundary conditions S, SS, ST, SST be extended *in the obvious way*. By the same arguments as in the last section we obtain the following theorem:

Theorem 4.6.1. *(The following assertions hold in the stated direction only; their converses are not in general valid.)*

(i) The model S,Q is ℓ_2- and GKS-unstable if Q is t-reversing.

(ii) The model SS,Q is ℓ_2- and GKS-unstable if Q is t-reversing or if Q is x_1-reversing and also x_j-reversing for at least one $j \geq 2$.

(iii) The model SST,Q is ℓ_2- and GKS-unstable if Q is x_1-reversing and/or t-reversing, and also x_j-reversing for at least one $j \geq 2$. ∎

Among the formulas Q considered by Abarbanel et al. are multidimensional versions of LF, CN, BE, and *MacCormack's scheme*. One sees readily that LF is t-reversing and x_j-reversing for each j, CN and BE are x_j-reversing for each j but not t-reversing, and MacCormack's scheme is not reversing in any variable. It turns out that all combinations of these schemes with S, SS, ST, or SST that are not ruled unstable by Thm. 4.6.1 admit no rightgoing solutions of any kind.

Abarbanel and Murman also consider a [illegible] model of (4.6.1), the ***Burstein scheme*** [Ab81]. Under this formula, a sawtoothed wave with $\kappa_x = \kappa_y = -1$ and $z = \pm 1$ turns out to have vector group velocity 0. The boundary formulas SS and SST support such a wave, which implies that the scheme is GKS-unstable, though not necessarily ℓ_2-unstable. Abarbanel and Murman do not report any experiments as to whether these modes give trouble in practice. See §5.4 for a discussion of this kind of borderline instability.

5. BORDERLINE CASES AND THE DEFINITION OF STABILITY

5.1 Introduction

In Chapter 4 we have seen that a difference model $\tilde{Q}$ of an initial boundary value problem may admit solutions exhibiting various degrees of instability. At one extreme, $\tilde{Q}$ may possess a ***strictly rightgoing eigensolution*** (i.e. $|z| > 1$ — a *Godunov-Ryabenkii eigensolution*), which grows exponentially with n and is therefore unstable in every reasonable measure (Thms. 4.2.1, 4.2.2). Or it may admit a ***strictly rightgoing generalized eigensolution*** (i.e. $|z| = |\kappa| = 1$ with positive group velocity) with an infinite reflection coefficient, as in §4.1, and we believe this situation is unambiguously unstable too. At the other extreme, $\tilde{Q}$ may be ***GKS stable,*** admitting no solution consisting of a combination of rightgoing modes of any kind (Thms. 4.3.1, 4.3.2). In this event it will behave stably in almost any sense. The complications come when one investigates situations between these two extremes, and this chapter is devoted to looking at some of these borderline cases. The guiding questions are, what is the meaning of stability for initial boundary value problems? How appropriate is the GKS stability definition?

We are mainly concerned with two classes of borderline cases. Suppose that $\tilde{Q}$ is GKS-unstable, admitting a rightgoing eigensolution or generalized eigensolution $z^n\phi$. Then how does $\tilde{Q}$ behave, if

(1) The reflection coefficient matrix $(D^{[r]})^{-1}D^{[\ell]}$ (3.6.5) corresponding to $z^n\phi$ is finite rather than infinite?

(2) $z^n\phi$ contains no *strictly* rightgoing modes? (i.e. $|z| = 1$, and for each mode in ϕ, either $|\kappa| < 1$ or $C = 0$?)

The various combinations implied by (1) and (2) do not exhaust the range of GKS instabilities, but we believe they touch the important issues. In this chapter our aim is to examine these problems, illustrating them with numerical experiments, in

order to demonstrate the complexity of the stability question for initial boundary value problems and to reach some tentative conclusions. Unfortunately, it has not been possible to be rigorous here, and our conclusions will be expressed as a series of "observations," not theorems. We do not attempt to state these observations precisely, and we do not claim that they hold as stated for all possible problems. What we do claim is that the observations capture some of the fundamental mechanisms that cause instability, and that many of them could probably be made rigorous, after appropriate modifications of details.

In §5.2 and §5.3 we consider situations (1) and (2), respectively. We will see that all of these borderline GKS-unstable situations behave stably in some respects. Section 5.4 describes the "transparent interface anomaly," a problem exhibiting both borderline features (1) and (2), which behaves stably in almost all respects and is in fact ℓ_2-stable. In §5.5 we summarize the main conclusions of this chapter, and of the dissertation, concerning borderline cases and the definition of stability for initial boundary value problems.

5.2 GKS-unstable solutions with finite reflection coefficients

For a general diagonalizable model $\tilde{Q}$ of a hyperbolic initial boundary value problem, we derived in §3.6 the equation

$$D^{[r]}a^{[r]} + D^{[\ell]}a^{[\ell]} = 0 \tag{5.2.1}$$

relating rightgoing and leftgoing modes at the boundary with uniform time dependence z^n. Here $a^{[r]}$ and $a^{[\ell]}$ are coefficient vectors of length n_r and n_ℓ, respectively, and $D^{[r]}$ and $D^{[\ell]}$ are matrices of dimension $n_r \times n_r$ and $n_r \times n_\ell$. According to Thm. 4.3.1, $\tilde{Q}$ is GKS-stable if and only if $D^{[r]}$ is nonsingular for all z with $|z| \geq 1$, in which case for any such z, $a^{[\ell]}$ determines $a^{[r]}$ by means of the formula

$$a^{[r]} = -(D^{[r]})^{-1}D^{[\ell]}a^{[\ell]}. \tag{5.2.2}$$

On the other hand if $D^{[r]}$ is singular for some $z = z_0$ with $|z_0| \geq 1$, then $(D^{[r]})^{-1}$ is undefined, and there is a risk that we may have in effect an infinite reflection coefficient.

What happens to (5.2.2) in this case? Obviously the equation as it stands has no meaning. However, assume that $D^{[r]}$ and $D^{[\ell]}$ are smooth functions of z in a point set Ω consisting of the intersection of $|z| \geq 1$ with a neighborhood of z_0, and that $D^{[r]}(z)$ is nonsingular in $\Omega - \{z_0\}$. The bases of right- and leftgoing solutions with respect to which $a^{[r]}$ and $a^{[\ell]}$ are defined also depend on z, but let us assume that this dependence is also smooth in Ω. Consider the limiting matrix

$$A_0 = \lim_{\substack{z \to z_0 \\ z \in \Omega}} \left(D^{[r]}(z)\right)^{-1} D^{[\ell]}(z). \tag{5.2.3}$$

The existence and behavior of A_0 will depend on whether $D^{[\ell]}(z_0)$ has singular behavior that cancels the singularity of $D^{[r]}(z_0)$. We consider three possibilities:

- If the product in (5.2.3) blows up as $z \to z_0$, then the limit does not exist, and A_0 is *infinite*.
- If the limit exists, then z_0 is a removable singularity, and A_0 is *finite*.
- Suppose that A_0 exists and is finite, and moreover

$$\ker(D^{[r]}(z_0)) \bigcap \mathrm{range}(A_0) = \{0\},$$

that is, if $D^{[r]}(z_0)a^{[r]} = 0$ with $a^{[r]} \neq 0$, then there exists no vector $a^{[\ell]}$ such that $a^{[r]} = A_0 a^{[\ell]}$. Then A_0 is *zero* (with respect to all unstable rightgoing solutions).

If we specialize the discussion to scalar problems with one leftgoing and one rightgoing solution for each z with $|z| \geq 1$, then (5.2.2) becomes

$$a^{[r]}(z) = -\frac{d^{[\ell]}(z)}{d^{[r]}(z)} a^{[\ell]}(z), \tag{5.2.4}$$

where each letter denotes a scalar. Section 3.2 derived many reflection coefficient functions of this kind. Q is GKS-unstable if and only if $d^{[r]}(z_0) = 0$ for some z_0 with $|z_0| \geq 1$. The limiting reflection coefficient will be infinite, finite, or zero depending on whether $d^{[\ell]}$ has a zero at $z = z_0$ of order lower than, equal to, or higher than that of $d^{[r]}$.

The question we wish to ask is: assuming Q is GKS-unstable, how is its unstable behavior, if any, affected by whether A_0 is infinite, finite, or zero?

Let $u_t = u_x$ be modeled by LF with $\lambda = \frac{1}{2}$ for $j \geq 1$. It turns out that by letting Q consist of various boundary formulas for v_0^{n+1} together with this scheme, we can cover a full range of degrees of stability. Consider the four possibilities $\alpha, \beta, \gamma, \delta$ listed in Table 5.1. We will now use these examples to explore the significance of reflection coefficients.

TABLE 5.1

Label	$v_0^{n+1} = \ldots$	Reflection function A	GKS-unstable mode $(\kappa_\ell, \kappa_0, z_0)$	$A_0 \equiv A(z_0)$	$C(\kappa_0, z_0)$
α	v_1^n	$-\frac{z-\kappa_\ell}{z-\kappa_0}$	GKS-stable	–	–
β	$v_0^n + \frac{\lambda}{2}(v_2^n - v_0^n)$	$-\frac{2(z-1)+\lambda(1-\kappa_\ell^2)}{2(z-1)+\lambda(1-\kappa_0^2)}$	$(1, -1, 1)$	0	$+1$
γ	$\frac{1}{2}(v_0^n + v_2^n)$	$-\frac{(2z-1)-\kappa_\ell^2}{(2z-1)-\kappa_0^2}$	$(1, -1, 1)$	$\frac{\lambda-1}{\lambda+1}$	$+1$
δ	v_1^{n+1}	$-\frac{1-\kappa_\ell}{1-\kappa_0}$	$(-1, 1, -1)$	∞	$+1$

Initial data

DEMONSTRATION 5.1. First let us consider stability with respect to the initial data $f(x)$. (We believe that the same ideas apply to stability with respect to forcing data $F(x,t)$.) Figure 5.1 shows a set of experiments in extension of the computation of Demo. 4.2. In every case, the LF model of $u_t = u_x$ with $\lambda = .5$ has been applied on $[0,1]$ for both $h = 1/50$ and $h = 1/100$. The initial distribution is the Gaussian

$$v_j^n = e^{-(x_j - \frac{1}{2})^2/.05^2} \kappa_\ell^j z_0^n \qquad \begin{array}{c} 0 \leq j \leq 1/h \\ n = 0, 1 \end{array} \tag{5.2.5}$$

with κ_ℓ chosen equal to the leftgoing wave value corresponding to the unstable rightgoing solution κ_0, i.e. $\kappa_\ell = 1$ for problems α, β and γ, $\kappa_\ell = -1$ for δ. Each pair of plots shows the initial data at $t = 0$ and the result at $t = 0.5$.

For the "standard" unstable case δ with $A_0 = \infty$, Fig. 5.1 shows a great growth in amplitude, as in Demo. 4.2. Obviously this is unstable in any reasonable sense. But for cases β and γ, just as for the GKS stable example α, no such growth is evident. We tentatively conclude:*

Observation 5.1. *Unstable amplification of initial data occurs only if an infinite reflection coefficient is present.*

However, even though no significant amplification takes place, a difference model may fail to converge as the mesh is refined. Example γ in Fig. 5.1 illustrates this. The smooth initial pulse ought to propagate across $x = 0$ and disappear, but instead, a reflected pulse is generated that evidently does not decrease in amplitude when h is

*Recall the disclaimer of §5.1 regarding these observations.

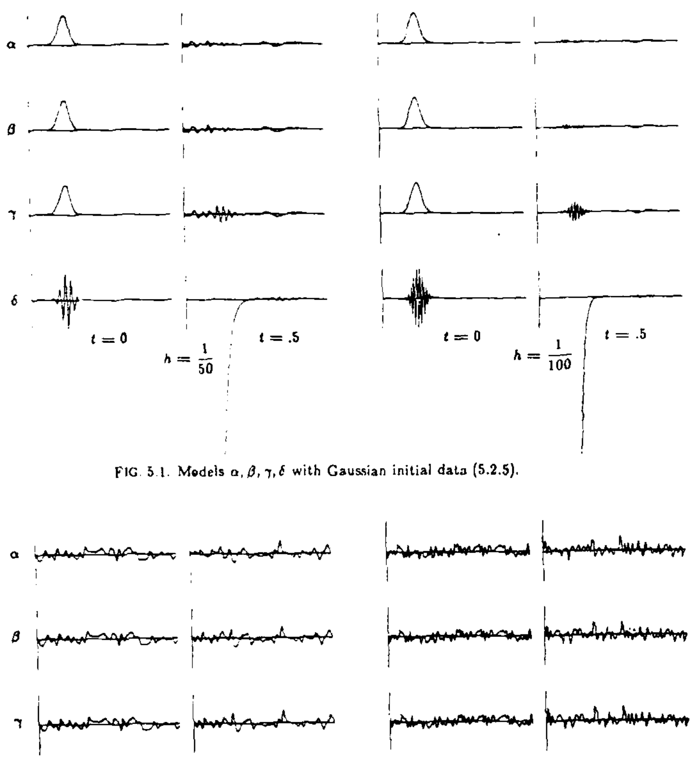


FIG. 5.1. Models $\alpha, \beta, \gamma, \delta$ with Gaussian initial data (5.2.5).

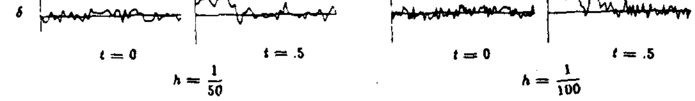


FIG. 5.2. Models $\alpha, \beta, \gamma, \delta$ with random initial data (5.2.6).

cut in half. In fact, it has amplitude very close to the limit $|A_0| = |\frac{\kappa-1}{\kappa+1}| = 1/3$ listed in Table 5.1. By contrast, the result for boundary condition β in Fig. 5.1 is virtually indistinguishable from the result for the GKS-stable case α, and it certainly appears that convergence is taking place. We propose

Observation 5.2. *Nonconvergence in a problem driven by smooth initial data occurs only if a nonzero reflection coefficient is present.*

DEMONSTRATION 5.2. One may wonder whether the same observations remain valid if a more complicated initial data distribution is considered. In Fig. 5.2, Demo. 5.1 is repeated with uniformly distributed *random* initial data,

$$v_j^n = \frac{1}{4}\,\text{random}_{[-1,1]}, \qquad \begin{matrix} 0 \le j \le 1/h \\ n = 0, 1 \end{matrix} \tag{5.2.6}$$

The plots show that the GKS-unstable problems β and γ are virtually indistinguishable from the GKS-stable problem α. But in the infinite reflection coefficient case δ, the computation is completely unstable. This supports Observation 5.1. This experiment does not shed any further light on Observation 5.2.

Boundary data

DEMONSTRATION 5.3. Now let us look at unstable behavior with respect to boundary data. In Fig. 5.3, Figs. 5.1 and 5.2 are duplicated with the new initial data distribution

$$v_1^0 = \frac{8}{5}, \qquad v_0^1 = \frac{4}{5}, \qquad v_1^1 = \frac{8}{15}, \tag{5.2.7}$$

which is the same as in Demo. 4.1 up to a scale factor. This amounts to an initial input of more or less random energy at the boundary. Fig. 5.3 shows that as t increases, spontaneous rightgoing waves are generated in all three cases β, γ, δ. Their amplitudes differ, but qualitatively all are the same (except of course for the difference in κ_0 between β-γ and δ). They are all qualitatively different from the GKS-stable problem α, where the initial data has apparently caused a rightgoing pulse of finite duration. A table of $\|v\|_2$ as a function of t confirms that a linear growth in energy is taking place in problems β–δ, but there is no growth for problem α. We conclude:

Observation 5.3. *A GKS-unstable difference model acts unstable with respect to boundary data regardless of whether the reflection coefficient is zero, finite, or infinite.*

This observation is in keeping with the fact that Thm. 4.2.1 made no mention of reflection coefficients.

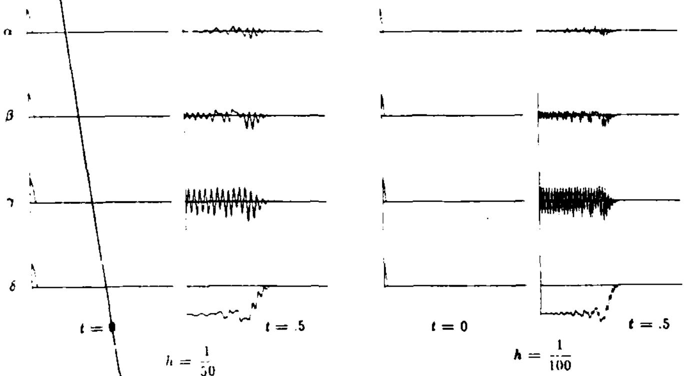


FIG 5.3. Models $\alpha, \beta, \gamma, \delta$ with three-point initial data **(5.2.7)**.

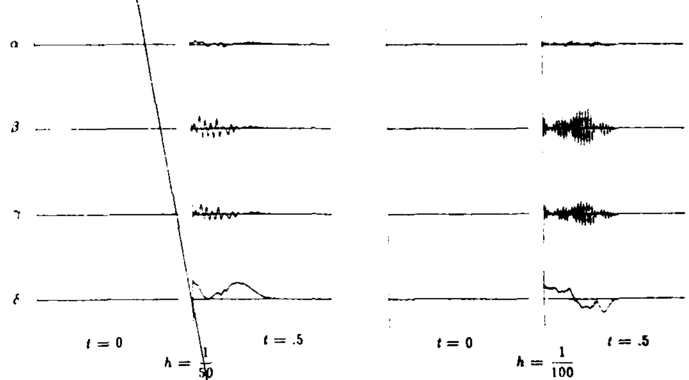


FIG 5.4. Models $\alpha, \beta, \gamma, \delta$ with random boundary data **(5.2.8)**.

DEMONSTRATION 5.4. Again, it is a good idea to confirm the observation with a randomized signal. Fig. 5.4 shows the usual collection of plots, except that now the forcing stimulus comes from inhomogeneous random boundary data,

$$g^n = \frac{1}{4}\,\mathrm{random}_{[-1,1]}. \tag{5.2.8}$$

As before, it is apparent that all three GKS-unstable boundary formulas β–δ lead to energy radiation into the interior. As t increases, the amplitude of the wave near the boundary will follow a kind of random walk, achieving amplitudes on the order of $\sqrt{n}$. By choosing g^n to be a regular wave at the unstable frequency ω_0, we could convert this to quite dramatic growth proportional to n. The GKS-stable problem α, on the other hand, behaves more like a damped random walk, with uniformly bounded amplitude, because in a stable problem energy input at the boundary moves directly into the interior, rather than remaining at the boundary to radiate energy forever. All of this gives further support to Observation 5.3.

Two boundaries

Suppose that in a problem involving two boundaries, say, a model of $u_t = u_x$ on the strip $x \in [0,1]$, $t \geq 0$, one or both of the boundary schemes is GKS-unstable because of a generalized eigensolution. We have seen that the unstable growth associated with such a boundary in isolation may be weak, but Kreiss points out in various papers that in the presence of the second boundary, the instability may be converted to catastrophic growth at an exponential rate [Kr71,Kr73,Gu72]. This phenomenon is one justification of the strictness of the GKS stability definition. In fact Kreiss shows the following [see Thm. 5.4 of [Gu72] for the precise formulation]:

Theorem 5.2.1 *[Gu72]. Let Q be a Cauchy stable model of a hyperbolic system on the strip $x \in [0,1]$, $t \geq 0$. Suppose that the initial boundary value problem obtained by removing the boundary at $x = 1$ to $x = +\infty$ is GKS stable, and so is the one obtained by removing the boundary at $x = 0$ to $x = -\infty$. Then Q is GKS stable.*

Proof. See §11 of [Gu72] and also §2 of [Kr66]. The basis of the argument is the invariance of GKS-stability with respect to perturbations of size $O(k)$; the effect of each boundary on the other can be shown to be of this order as $k \to 0$. ∎

DEMONSTRATION 5.5. The theorem makes no mention of reflection coefficients. However, observe what happens when an experiment of this kind is tried on the examples α–δ we have been studying. Table 5.2 summarizes the results of an experiment identical to that of Demo. 5.3, but carried up to $t = 14$, which is time enough for many reflections between the boundaries to take place. Each entry shows the ℓ_2 norm $\|v^n\|_2$ at a fixed time step:

	α	β	γ	δ
$t = 0$	.154	.154	.154	.154
1	.150	.167	.170	.513
2	.104	.132	.136	.990
6	.083	.124	.135	5.66×10^3
10	.051	.118	.149	1.80×10^7
14	.073	.130	.174	4.52×10^{10}

TABLE 5.2

two boundaries
LF $h = \frac{1}{50}$ $\lambda = \frac{1}{2}$

Again the equation $u_t = u_x$ was modeled on $[0,1]$ by LF with $h = 1/50$, and the boundary condition at $x = 1$ was $v_{50}^{n+1} = 0$. The table shows that problem δ exhibits catastrophic growth, but for problems α, β, γ there is no growth at all. Obviously the GKS-stable problem α has no advantages here over the GKS-unstable problems β and γ. We propose:

Observation 5.4. *An unstable generalized eigensolution can cause exponential growth when a second boundary is introduced only if the associated reflection coefficient is infinite.*

There is a simple argument involving z, κ, and $A(z)$ that explains why Observation 5.4 should hold. For this see §6.4 and §6.5, where we discuss two-boundary problems in detail.

5.3 GKS-unstable solutions with no strictly rightgoing components

Suppose that $\hat{Q}$, a difference model of an initial boundary value problem, admits an eigensolution or generalized eigensolution (4.2.6)

$$v_j^n = z^n \sum a_i \kappa_i^j \psi_i, \qquad a_i \neq 0 \tag{5.3.1}$$

with $|z| = 1$. (For simplicity we ignore defective modes.) The assumption $|z| = 1$ rules out Godunov-Ryabenkii eigensolutions, but the solutions that remain are GKS-unstable by Thm. 4.3.1. They fall into three categories, which correspond to positions (8), (5), and (7) of Table 2.1, respectively:

"Case $C > 0$:" For at a least one i, $|\kappa_i| = 1$ and $C_i > 0$.

"Case $C = 0$:" Not Case $C > 0$, but for at least one i, $|\kappa_i| = 1$.

"Case $|\kappa| < 1$:" Neither of above, i.e. $|\kappa_i| < 1$ for all i.

By definition, each signal $z^n \kappa_i \psi_i$ in (5.3.1) is rightgoing, but in the cases $C = 0$ and $|\kappa| < 1$, none of them are strictly rightgoing (§2.3). We want to investigate how this affects their unstable behavior, if any.

As in the last section, we will work with representative examples. Here is a contrived but very simple model of type $C = 0$:

$$\epsilon: \quad \text{LF for } u_t = u_x \text{ with } \lambda = \frac{1}{2}; \quad v_0^{n+1} = v_1^{n-2}. \tag{5.3.2}$$

(We continue as in the last section to label examples with Greek letters.) It is easy to verify that (5.3.2) admits the GKS-unstable generalized eigensolution $(\kappa, z) = (\pm i, \pm i^{1/3})$, for which one has $C = 0$.

For an example of type $|\kappa| < 1$ we turn to a dissipative Lax-Wendroff model:

$$\zeta: \quad \text{LW for } u_t = u_x \text{ with } \lambda = \frac{1}{3}; \quad v_0^{n+1} = 2v_2^{n+1} - v_1^{n+1}. \tag{5.3.3}$$

One readily verifies that this model admits the GKS-unstable eigensolution $(\kappa, z) = (-\frac{1}{2}, 1)$.

By straightforward computations of the sort we have done many times, one can see that examples ϵ and ζ share the feature that their right/left reflection coefficients are finite. (In fact one gets $A_0 = -1$ and $A_0 = 4$, respectively.) This will make it difficult to separate the effects of one borderline circumstance from those of the other. To get an example with $C = 0$ but $A_0 = \infty$, we invent the following 2×2 problem:

$$\eta: \quad \text{LF for } \begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x \text{ with } \lambda = \frac{1}{2}; \qquad u_0^{n+1} + v_0^{n+1} = u_1^{n-2} + v_1^{n-2}, \qquad v_0^{n+1} = v_1^n. \tag{5.3.4}$$

Like ϵ, problem η admits a rightgoing solution of type $C = 0$, namely $(\kappa, z) = (\pm i, \pm i^{1/3})$, $\psi = (1\ 0)^T$. But now the reflection coefficient with respect to (strictly) leftgoing energy incident in the v component is infinite. Let κ, μ be the κ variables for the u and v components, respectively. Then (5.2.1) takes the form

$$\begin{bmatrix} z^3 - \kappa & z^3 - \mu \\ 0 & z - \mu \end{bmatrix} \begin{pmatrix} a_1^{[r]} \\ a_2^{[r]} \end{pmatrix} + \begin{bmatrix} z^3 + 1/\kappa & z^3 + 1/\mu \\ 0 & z + 1/\mu \end{bmatrix} \begin{pmatrix} a_1^{[\ell]} \\ a_2^{[\ell]} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

and multiplying through by the inverse of the first matrix gives for (5.2.2), after some simplifications,

$$\begin{pmatrix} a_1^{[r]} \\ a_2^{[r]} \end{pmatrix} = -\begin{bmatrix} \frac{z^3+1/\kappa}{z^3-\kappa} & \frac{(z^3+z)(1/\mu-\mu)}{(z^3-\kappa)(z-\mu)} \\ 0 & \frac{z+1/\mu}{z-\mu} \end{bmatrix} \begin{pmatrix} a_1^{[\ell]} \\ a_2^{[\ell]} \end{pmatrix}. \tag{5.3.5}$$

For $(\kappa, z) \doteq (\pm i, \pm i^{1/3})$, the diagonal elements of this matrix are finite, but the upper-right element is infinite. Therefore we expect leftgoing energy in v with $\omega k = \pi/6$, hence $\xi h = \sin^{-1}(\sin\frac{\pi}{6}/\frac{3\lambda}{2}) = \sin^{-1}(\frac{2}{3})$, to stimulate a large response in u.

DEMONSTRATION 5.6. As a first test of examples ϵ–η, Figs. 5.5–5.7 repeat the computations of Demos. 5.2–5.4 (Figs. 5.2–5.4). The three figures show the response of models ϵ, ς, and η to the stimuli

Fig. 5.5: random initial data (5.2.6),

Fig. 5.6: random boundary data (5.2.8),

Fig. 5.7: three-point initial/boundary data (5.2.7).

For problem η, the forcing data are applied to v but not u, and both u and v are plotted, with the labels η_u and η_v. Since the v component of this problem is identical to problem α of the last section, except for the coefficient 3/2 in place of 1, the η_v plot gives a convenient GKS-stable comparison to the others. As before, each plot shows $t = 0$ and $t = .5$ for $h = 1/50$ and $h = 1/100$.

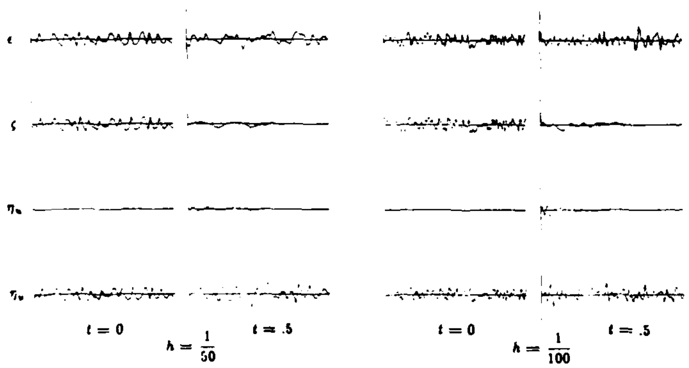


FIG. 5.5. Models $\epsilon, \varsigma, \eta$ with random initial data (5.2.6).

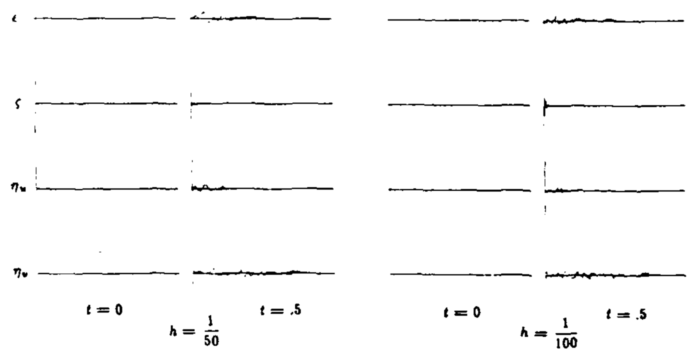


FIG. 5.6. Models $\epsilon, \varsigma, \eta$ with random boundary data (5.2.8).

FIG. 5.7. Models $\epsilon, \varsigma, \eta$ with three-point initial data (5.2.7).

Growing modes

The first thing to observe in these figures is that, in contrast to the situation with examples β–δ in Figs. 5.1–5.4, no catastrophic growth is taking place. For problem ζ, with $|\kappa| < 1$, all of the solutions shown are quite small, as the dissipativity would make one expect. Note that in Figs. 5.6 and 5.7, the distribution at $t = .5$ for this problem looks exactly like the eigensolution $(-\frac{1}{2})^j$. For problem ϵ, with $C = 0$ and $A_0 < \infty$, the random initial and boundary data do not seem to have caused instability (compare problem α in Figs. 5.2 and 5.4), but the situation with the three-point initial data at the boundary is not so clear. In fact near the boundary in Fig. 5.7, the solution looks approximately like the "$4h$ wave" with $\kappa = \pm i$ that the GKS theory says is unstable. The results for the u component in problem η, with $C = 0$ but $A_0 = \infty$, are similar, while the v component is entirely stable, which is what one expects from (5.3.4) or (5.3.5).

The situation is clarified if we look at ℓ_2 norms as a function of t for the three-point problems of Fig. 5.7. Table 5.3 lists $\|v^n\|_2$ for ϵ–η for $h = 1/50$ and $h = 1/100$ at times $t = nk = 0, .2, \ldots, 1$.

	ϵ	ζ	η_u	η_v	
$t = 0$	.135	.135	0	.135	
.2	.132	.125	.104	.077	
.4	.114	.125	.076	.075	$h = \frac{1}{50}$
.6	.099	.125	.096	.075	
.8	.117	.125	.093	.074	
1.0	.110	.125	.076	.070	TABLE 5.3
$t = 0$	.096	.096	0	.096	3-point boundary data
.2	.081	.089	.054	.054	
.4	.083	.089	.066	.054	
.6	.072	.089	.066	.054	
.8	.077	.089	.058	.057	$h = \frac{1}{100}$
1.0	.080	.089	.062	.053	

In no case do we observe any growth in energy. (Note how the numbers confirm that for problem ζ, the solution rapidly settles down to the form of a fixed eigensolution.) Therefore we suggest:

Observation 5.5. *In case $C = 0$ or $|\kappa| < 1$, Q admits no solutions that grow steadily with t.*

Usually a difference scheme exhibits conspicuous instability in practical examples only if there is such a growing mode, so Obs. 5.5 explains why no instabilities are evident, for example, in Fig. 5.5. Of course we have long ago observed that non-strictly rightgoing solutions have zero energy flux (§3.3), and one would expect Obs. 5.5 to hold as a consequence of this.

Initial data

The absence of growing modes, even if we could prove it rigorously, would not imply ℓ_2-stability, because there could still exist initial data distributions that would grow arbitrarily much at first before ultimately leveling off. Nevertheless, we conjecture that problem ϵ is ℓ_2-stable, as defined in §4.2, despite being GKS-unstable. If true, this can probably be proved by an energy method argument [Ri67], and possibly also by an application of the ideas of §3.5. In general, one appears to have something like the following:

Observation 5.6. *In case $C = 0$ or $|\kappa| < 1$, Q is stable with respect to initial data provided that the reflection coefficients are finite.*

DEMONSTRATION 5.7. By contrast, a problem with $A_0 = \infty$ need not be stable with respect to initial data. To demonstrate this, Fig. 5.8 shows a computation with problem η in which initial data have been chosen to stimulate as much growth as possible, as was done for examples β–δ in Fig. 5.1. The initial data are

$$\begin{aligned} v_j^0 = v_j^1 &= \cos(\xi x)\, e^{-[(x-.25)/.1]^2} \qquad (x = jh), \\ u_j^0 \equiv u_j^1 &\equiv 0, \end{aligned} \tag{5.3.6}$$

with $\xi h = \sin^{-1}(\frac{2}{3})$. Fig. 5.8 shows the resulting signals u and v at $t = 0, .5, 1$ for $h = 1/100$ and $h = 1/100$. It appears that there is some instability, but it is extremely weak. The initial wave (5.3.6) with $h = 1/100$ has eight times as many grid points in the wave packet as (4.1.6) with $h = 1/200$ or (5.2.5) with $h = 1/100$, yet it generates nothing like the 18-fold amplitude increase that we see in Fig. 4.2a,b and that is lurking off-scale in Fig. 5.1. Moreover, the signal that it generates does not radiate continually from the boundary, but evidently loses amplitude as it drifts into the interior.

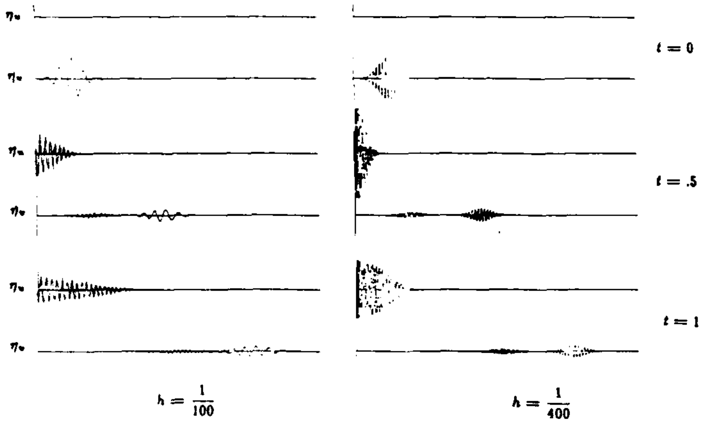


FIG. 5.8. Unstable reflection with $C = 0$ in model η. The initial packet is the Gaussian (5.3.6).

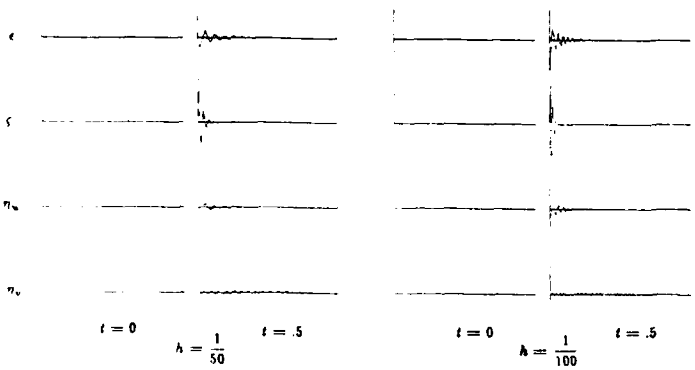


FIG. 5.9. Models ϵ, ζ, η with periodic boundary data (5.3.7).

Again, a table of ℓ_2 norms makes it clear what is going on. Table 5.4 shows $\|u\|_2$ and $\|v\|_2$ in this problem for both values of h and various times.

	$h = \frac{1}{100}$		$h = \frac{1}{400}$	
	η_u	η_v	η_u	η_v
$t = 0$	0	.177	0	.177
.2	.165	.150	.207	.143
.4	.198	.060	.301	.059
.6	.185	.060	.293	.059
.8	.182	.060	.291	.059
1.0	.184	.060	.292	.059

TABLE 5.4
"bad" initial conditions

For $h = 1/100$, we observe an amplification $\|u(1)\|/\|v(0)\|$ of about 1, and for $h = 1/400$ it has increased to more like 2. Evidently the ratio can be made arbitrarily large by refining the mesh. But it is hardly large as things stand, and - confirming Obs. 5.5 - there are no solutions in evidence that grow with t. We propose:

Observation 5.7. *If $\tilde{Q}$ has $C = 0$ but $A = \infty$, it is weakly unstable with respect to initial data.*

Boundary data

From Figs. 5.6 and 5.7, we expect that if $\tilde{Q}$ has $C = 0$ or $|\kappa| < 1$, then it will not be dramatically unstable with respect to boundary data. In fact, as in Obs. 5.7, it turns out that there is a weak instability. As we found in Obs. 5.3, the presence of this instability does not depend on whether A_0 is infinite.

DEMONSTRATION 5.8. Fig. 5.9 shows an experiment like those of Figs. 5.6 or 5.8, except that now the computation is forced by regular boundary data oscillating at the GKS-unstable frequency. The boundary condition is

$$v_0^{n+1} = (\text{homog. b.c.}) + \frac{1}{10}\cos \omega k n \tag{5.3.7}$$

with $\omega k = \pi/6$, 0, and $\pi/6$ for ϵ, ζ, and η, respectively, and the figure shows $h = 1/50$ and $h = 1/100$. The results are much like those of Fig. 5.7, but stronger. Some instability is definitely in evidence for all three problems (note the small amplitude of the forcing term in (5.3.7)), and it grows stronger as h is refined or as t increases with the boundary function left on. Table 5.5 confirms this with a record of ℓ_2 norms, which for the case of problem ζ become fairly large. For comparison with these numbers, the norm of the forcing function in (5.3.7) is approximately $\sqrt{t}/10$.

	ϵ	ζ	η_u	η_v	
$t = 0$	0	0	0	0	
.2	.039	.112	.019	.023	
.4	.059	.220	.037	.031	$(h = \frac{1}{50})$
.6	.092	.327	.054	.039	
.8	.102	.435	.059	.044	
1.0	.122	.543	.076	.049	TABLE 5.5
$t = 0$	0	0	0	0	"bad" boundary data
.2	.042	.156	.026	.022	
.4	.073	.309	.042	.032	
.6	.107	.462	.064	.039	
.8	.124	.615	.078	.044	$(h = \frac{1}{100})$
1.0	.146	.769	.087	.050	

Nevertheless, to achieve this amount of growth we had to stimulate just the right frequency, and if we had done the same for the strictly rightgoing problems of the last section, the result would have been much more dramatic. We conclude:

Observation 5.8. *A model with $C = 0$ or $|\kappa| < 1$ is weakly unstable with respect to boundary data regardless of whether the reflection coefficient is zero, finite, or infinite.*

Despite the impressive ℓ_2 norm, it is obvious (and expected) that nothing happens in case ζ except at the boundary:

Observation 5.9. *In a problem of type $|\kappa| < 1$, any unstable growth is confined to the region near the boundary.*

Two boundaries

Observation 5.9 suggests that in a two-boundary problem, as we considered in the last section, the presence of an unstable boundary of type $|\kappa| < 1$ probably will not cause exponential growth. Indeed this is true, as our arguments of §6.5 will show. We have the following complement of Obs. 5.4:

Observation 5.10. *An unstable boundary can cause exponential growth in a two-boundary model if it is of type $C = 0$, but not if it is of type $|\kappa| < 1$.*

DEMONSTRATION 5.9. To illustrate Obs. 5.10 experimentally, we ran problems

ϵ, ζ, η up to $t = 400$ with random initial data (5.2.6), in each case on a grid with $h = 1/50$. The boundary conditions at the right hand side were $v_{50}^{n+1} = 0$ for problems ϵ and ζ, and

$$u_{50}^{n+1} = 0, \qquad u_{50}^{n+1} - v_{50}^{n+1} = u_{49}^{n} - v_{49}^{n}$$

for problem η. (The complexity of the latter formula is needed to introduce some coupling between u and v at the right boundary.) The results are summarized in Table 5.6.

	ϵ	ζ	η_u	η_v	
$t = 0$	.153	0	.153	.153	
100	.022	.037	18.2	3.42	TABLE 5.6
200	.021	.037	2.41×10^4	3.65×10^3	two boundaries
300	.021	.037	3.18×10^7	4.78×10^6	
400	.018	.037	4.15×10^{10}	6.38×10^9	

Evidently ϵ and ζ generate no growth at all, while for the η model there is exponential growth, but it is much weaker than in Table 5.2e (note the large values of t in the table). These results are in keeping with Observations 5.4 and 5.10.

5.4 The transparent interface anomaly: inflow-outlow theorems

The "transparent interface anomaly" is an example that in addition to being rather startling, proves that a difference model may be GKS-unstable but at the same time ℓ_2-stable. Consider any of the mesh-refinement problems of Example 3.4 (§3.2), in which a grid with $h = h_-$ for $j < 0$ is connected to a grid with $h = h_+$ for $j > 0$, and let LF be *the formula applied on each side*. If $h_- \neq h_+$, then all three interface formulas considered in §3.2 are stable, except for the "coarse-mesh" formula in the case when h_+/h_- is an even integer (Thm. 4.5.4). But now, consider the degenerate situation $h_- = h_+$. In all three cases, for $h_- = h_+$ the interface conditions become equivalent to the LF formula at $j = 0$, and one is left with LF applied at all points $j \in Z$. Since LF is Cauchy stable, the result must be ℓ_2-stable. Nevertheless, this degenerate case of a *"transparent interface"* is GKS-unstable. This fact appears not to have been pointed out before, even though [Br73], at least, studies GKS stability for mesh refinement schemes.

It is easy to see why the transparent interface model is unstable according to the GKS definition. The GKS-unstable solution that it admits is a generalized eigensolution with $|z| = 1$ and $\kappa_0 = \pm 1$ on both sides of the interface. By Fig. 1.1a or (1.2.5), this wave has group velocity $C = 0$, and this implies that if a wave packet of this kind is located initially around $j = 0$, it will remain approximately fixed there as t increases. We have already observed in §4.3 that because of the boundary integral in the definition of GKS-stability, such stationary behavior constitutes GKS-instability.

In terms of Thm. 4.3.1, the instability results from the fact that the wave with $C = 0$ is by definition both rightgoing *and* leftgoing (position (5) in Table 2.1), and so one may think of it as rightgoing for $j > 0$ and leftgoing for $j < 0$. Or in terms of Thm. 4.3.2, the instability is due to the fact that for $\kappa_i = \kappa_\ell = \pm 1$, the denominators of the reflection coefficient functions (3.2.20), (3.2.21), (3.2.27) vanish.

We have shown:

Observation 5.11. *GKS instability does not imply ℓ_2 instability.*

In this instance it is possible to state the conclusion as a theorem, which we have generalized in the obvious way:

Theorem 5.4.1 *Let Q be any scalar or vector difference model applied for $j \in Z$ that admits a steady state solution $z^n\kappa^j v$ with $|z| = |\kappa| = 1$ and $C(\kappa, z) = 0$. Let $j = j_0$ be thought of as an interface point of Q and the GKS theory applied by "folding" the model at this point (§3.6). Then the result is GKS unstable.*

Proof. GKS-instability follows from Thm. 4.3.1, since $z^n\kappa^j v$ can be thought of as rightgoing on the right and leftgoing on the left. ∎

The transparent interface problem as described above involves a doubly borderline kind of instability. First, the group velocity is not positive, but 0. Second, the reflection coefficient must have the value 0 too—for no energy can be reflected at an interface that is equivalent to the absence of an interface. Thus we are dealing with the intersection of the weakly unstable classes of the last two sections, and in Observations 5.2 and 5.6 we have already proposed that such a problem will be stable in ℓ_2 norm. In fact, in this problem we have not only $\Lambda_0 = 0$, but $\Lambda(z) = 0$ for all z. These considerations confirm that no reflection mechanisms are operative to make the transparent interface model act unstable with respect to initial data. Of course it would be weakly unstable with respect to forcing data at the point $j = j_0$, according to Obs. 5.5, but if the interface comes from a mesh refinement strategy, one would never apply such data.

All together, it is clear that the GKS assessment of instability is rather misleading for the transparent interface problem, and that this is due to the fact that the GKS theory is oriented towards boundaries while the transparent interface problem concerns initial data. It is natural to suppose that there may be other models of initial boundary value problems for which the GKS result is also unreasonable, but where the true state of affairs would not be so obvious.

• • •

A "strict transparent interface anomaly"

The identically zero reflection coefficient function is the essential reason for the stability of the transparent interface anomaly, not the fact that the GKS-unstable generalized eigensolution involved has $C = 0$. To see this, consider now the same setup as before, but with LF replaced by LF^2, the Leap Frog model of $u_{tt} = u_{xx}$ (§3.2, App. A). For LF^2, there are two distinct waves with $\kappa = z = 1$, one strictly rightgoing with $C = 1$, and one strictly leftgoing with $C = -1$. The dispersion plot of App. A makes this clear. As a result, LF^2 must admit solutions of the kind illustrated in Fig. 5.10, which are *strictly outgoing* from both sides of a transparent interface:

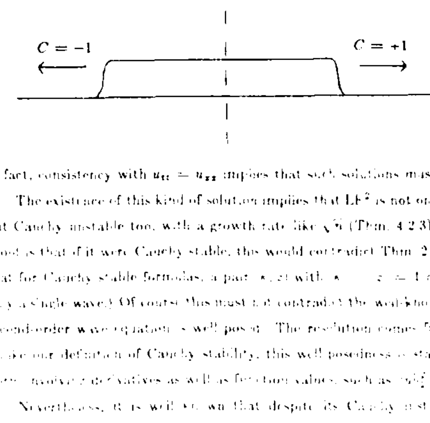


FIG. 5.10

In fact, consistency with $u_{tt} = u_{xx}$ implies that such solutions must be possible.

The existence of this kind of solution implies that LF^2 is not only GKS unstable, but Cauchy unstable too, with a growth rate like $\sqrt{n}$ (Thm. 4.2.3). (An alternative proof is that if it were Cauchy stable, this would contradict Thm. 2.3.1, which states that for Cauchy stable formulas, a pair κ, z with $|\kappa| = |z| = 1$ can correspond to only a single wave.) Of course this must not contradict the well-known result that the second-order wave equation is well posed. The resolution comes from the fact that unlike our definition of Cauchy stability, this well posedness is stated in terms of a norm involving derivatives as well as function values, such as $\|u\|_2^2 + \|u_x\|_2^2$.

Nevertheless, it is well known that despite its Cauchy instability, LF^2 is in practice a completely reliable difference model. The standard way to show this, in analogy with the situation for $u_{tt} = u_{xx}$, would be to prove that LF^2 satisfies a bound involving function values and their first-order differences. But in the present context, we can view its stability in practice as a consequence of the identically zero reflection coefficients. That is, the LF^2 model possesses unstable resonant modes, but they are not significantly excited by initial data.

• • •

Inflow-outflow theorems

However, we will now show that the GKS-instability of the transparent interface problem, even in the case $C = 0$, is not completely irrelevant to questions of stability with respect to initial data. The following is a paraphrase of an "inflow-outflow" theorem, published first by Goldberg and Tadmor; for details see Thm. 2.1 of [Go81]. By "inflow" and "outflow" variables, we mean variables corresponding to components of the partial differential equation with characteristics pointing into or out of the region at the boundary, respectively — not to rightgoing or leftgoing modes admitted by the difference model. This use of the terms is standard.*

Theorem 5.4.2 *[Go81]. Let $\tilde{Q}$ be a diagonalizable Cauchy stable difference model consistent with a well posed hyperbolic initial boundary value problem. Assume that at the boundary the inflow variables, u, are given as functions of the outflow variables, v. Then Q is GKS-stable if and only if its restriction to the outflow variables v is GKS-stable.*

Sketch of proof. If the restriction of $\tilde{Q}$ to v is GKS-stable, then by the definition of GKS-stability (4.3.1), the t-integral of v at the boundary $x = 0$ can be estimated. These boundary values are just the boundary data for u, so it follows that u can be estimated too. ∎

We claim the following:

Observation 5.12. *An inflow-outflow theorem like Thm. 5.4.2 ceases to hold if "GKS-stable" is replaced by "ℓ_2-stable".*

*However, in keeping with our general policy of discussing stability with minimal reference to the properties of the differential equation, we observe that the same theorem holds for any splitting of a difference model into variables u and v.

To show this, imagine a problem with one outflow variable v and two inflow variables, u_1 and u_2. Let v and u_1 correspond to a folded transparent interface with $C = 0$ so that data in these components at some frequency ω_0 can remain stationary at the boundary. Let u_2 obey any model under which boundary data at frequency ω_0 generate a rightward flux of energy. Then obviously the system cannot be ℓ_2-stable, but its restriction to v is.

The difficulty again has to do with boundaries. The inflow-outflow idea depends on having control of the numerical solution along the boundary, since that is where the inflow and outflow variables are coupled. The GKS stability definition is strong enough for this to go through, because of the boundary integral it includes, while the definition of ℓ_2-stability is not. This would appear to be a point in favor of the GKS stability definition. However, since the transparent interface problem does not act unstable in any more usual sense, we are inclined to think that inflow-outflow theorems are probably too much to demand of a stability definition.

It is worth pointing out that not only can energy with $C = 0$ remain approximately fixed in space for a long time, but it can in principle accumulate in a spike at one point, causing large growth in the L_∞ norm. An example of this kind is given in §5 of [Tr82]. Thus in a GKS-unstable problem with ℓ_2-stable outflow components, it is possible that a small outflowing signal could generate a strongly unstable inflowing one that was large in amplitude, not just total energy.

5.5 Summary and discussion

For the Cauchy problem with constant coefficients, there are two possible mechanisms of instability: amplification factors z of modulus greater than unity, and defective amplification factors of modulus equal to unity (Thm. 2.2.2). If the coefficients are allowed to vary with x, t, or k, new possibilities arise [Ri67]. When a boundary is introduced, one must begin to consider Godunov-Ryabenkii eigensolutions, spontaneous radiation of parasitic waves, borderline waves with $C = 0$, trapped signals with $|\kappa| < 1$, and radiative solutions of defective type (Thms. 4.2.1, 4.2.3, 4.3.1). Adding a second boundary raises the prospect of reflections back and forth between the two boundaries, which we will see in §6.5 is itself complicated. Finally, the world of instabilities for nonlinear problems is largely uncharted. Altogether, many qualitatively different mechanisms can cause instability, and they range from the explosive to the nearly invisible. Nor can one assume that these various mechanisms will not interact to produce further complications.

The variety of stability questions that one may wish to answer is equally complicated. One may be concerned with a difference model driven by initial data, forcing data, or boundary data, or some combination of these, and one may or may not be willing to assume that they have some degree of smoothness. One may be interested in ℓ_2 or in maximum errors, at fixed time steps or averaged over time, in the field only or at the boundary also. One may want a guarantee that stability will be preserved when a second boundary is introduced, or when one outflow model is used to drive a distinct inflow model, or when undifferentiated terms or other perturbations are added. And of course, technical limitations inevitably lead to the consideration of further stability definitions that would never come up naturally, as one tries to find a workable compromise between what can be proved and what can be used.

In summary, the first point that *we wish to emphasize is this:*

> *Instability for difference models is caused by identifiable physical mechanisms, especially phenomena of dispersive wave propagation. A complete understanding of instability requires a recognition of these mechanisms. Different mechanisms are relevant to different stability questions. No single definition of stability, or identification of its cause, can satisfactorily account for all possibilities.*

We have reached many more specific conclusions about what physics normally causes what kinds of instability. The most important ones can be summarized as follows:

> *Instability with respect to initial data is usually associated with the existence of infinite reflection coefficients (Obs. 5.1, 5.2, 5.6, 5.7). Instability with respect to boundary data is usually associated with the existence of spontaneous strictly rightgoing solutions (Obs. 5.3, 5.5, 5.8, 5.9). Instability with respect to the introduction of a second boundary is associated with infinite reflection coefficients involving wavelike modes (Obs. 5.4, 5.10).*

The GKS theory represents an extreme point in several respects. For one thing, it goes far in the direction of emphasizing mathematical unity at the expense of naturalness, combining all stability issues into a single remarkably complicated definition, about which a remarkably simple theorem can be proved. Second, the GKS stability

definition is also close to extreme in its conservativeness: if a problem is GKS-stable, it is almost certainly stable in practice, whereas we have seen in this chapter that the converse does not hold. However, in some respects the GKS definition is not so strict. Its generous allowance for exponentially growing solutions leads to problems related to "P-stability" that we will discuss in §6.4; and its failure to give estimates at fixed time steps rather than integrated over all t makes its application to adaptive mesh refinement problems difficult (Joseph Oliger, private communication).

The GKS theory is focused on *boundaries.* The stability definition (4.3.1) requires that the solution along the boundary satisfy an estimate in terms of the data along the boundary, and the proof of the GKS theorem (which we have not discussed) gives evidence of this bias: it proceeds by reducing the difference model to a recurrence relation in j, with the boundary conditions for initial data, and the forcing data F are introduced only as an inhomogeneous term in this recurrence relation. In fact, the result labeled "main theorem" in the GKS paper is not our Thm. 4.3.1, but an assertion that GKS-stability is equivalent to a boundary estimate *(Thm. 5.1 of [Gu72])*. Initial data do not figure naturally in the theory at all, and must be introduced by way of the forcing function F at the cost of a factor of h (Thm. 3.1 of [Gu72]), or by way of the boundary data g at the cost of a smoothness restriction (Thm. 2.1 of [Gu81]). Ideally, an analogous theory would be available that was fundamentally oriented towards initial conditions instead, but although Osher's results of [Os69b] are of this type, they do not have full generality.

Our summary assessment of the GKS theory is this:

> *There is probably no better all-purpose stability criterion than the GKS determinant condition. However, the theory in support of this condition, in particular the GKS stability definition, are relatively unsatisfactory, and fully justify the determinant condition only with respect to the problem of estimating boundary values in terms of boundary data. For additional insight in particular problems, it is worth checking whether any GKS-unstable solution has infinite reflection coefficients and strictly rightgoing modes.*

6. STABILITY FOR MODELS WITH SEVERAL BOUNDARIES OR INTERFACES

6.1 Introduction

In this final chapter we consider difference models containing two or more boundaries or interfaces. The question is, when is such a model stable? In most cases the GKS theory gives a procedure for answering this question, but the algebra involved is often very complicated, and in addition regrettably problem-specific. To avoid these difficulties, it is natural to look for stability results that depend only on the properties of each interface independently. One asks, *what properties of an interface can guarantee that models containing several such interfaces will be stable, or unstable?*

The simplest problem of this kind is that of modeling a hyperbolic system of equations on a strip, say $0 \le x \le 1$, with numerical conditions prescribed on each boundary. For this the GKS theory gives what appears to be the ideal result, which we quoted as Thm. 5.2.1: for GKS-stability of the strip model, GKS-stability of each *boundary* individually is sufficient. In fact, Thm. 5.2.1 is not quite ideal. The difficulty is that for fixed h and k, a GKS-stable difference model often exhibits exponential growth for a problem whose solution does not grow, and the rate of growth need not decrease as h and k are reduced unless the model is totally dissipative. We will look at this problem in §6.4. Still, for most purposes Thm. 5.2.1 is good enough for realistic strip problems.

We will be mainly interested in a different class of difference models, those in which the interfaces are not separated by a fixed distance Δx in x as $h, k \to 0$, but by a fixed number of mesh intervals Δj. (Of course in reality, every computation is done on a finite mesh, so this distinction is sometimes delicate.) Such problems come up, for example, when one has an *initial boundary value problem* model that involves two or more distinct boundary formulas in addition to the interior formula, as would normally be used when the interior scheme has order of accuracy greater than two [O171,O176]. It is not standard to view such composite schemes as consisting of fixed formulas separated by interfaces, but we believe this approach may be useful. They might also occur in modeling an *adaptive mesh refinement scheme*, where there can be no guarantee that one interface between meshes will remain a fixed distance from the next as h is decreased.

For these multi-interface problems of "fixed Δj type", no theorem as simple as Thm. 5.2.1 holds, and we will demonstrate this in §6.3. However, §6.4 and §6.5 will show that stability results can sometimes be obtained by arguments based on reflection coefficients.

6.2 One interface: results of Ciment and Tadmor

INTERFACE PROBLEM. Consider a scalar difference model $\bar{Q}$ consisting of one formula Q_- applied for $-\infty < j < j_0$ coupled with a second formula Q_+ applied for $j_0 \le j < \infty$. This is an interface of the "abrupt change" type considered in §3. $\bar{Q}$ may represent a discontinuous physical system, or the interface may be a numerical one (mesh refinement, hybridization). Assume that Q_- and Q_+ each satisfy Ass. 3.1, so that Q_- has stencil parameters ℓ_-, r_- and admits exactly r_- leftgoing and ℓ_- rightgoing solutions for all z with $|z| \ge 1$, similarly for Q_+. We have discussed the stability of such problems in §3–§5. In Thm. 4.3.1 we saw that $\bar{Q}$ is GKS-unstable if and only if it admits some steady-state solution that is outgoing from both sides of the interface, as suggested by Fig. 6.1:

FIG. 6.1

Each arrow in the figure represents one signal with energy flux in the indicated direction (positions (1)–(5) in Table 2.1 on the left, (5)–(9) on the right). To prove stability, one must show that the kind of configuration illustrated cannot occur.

INITIAL BOUNDARY VALUE PROBLEM. Alternatively, consider a similar scalar model $\bar{Q}$ for an initial boundary value problem. For $j \ge j_0$, $\bar{Q}$ consists of a fixed formula Q_+ satisfying Ass. 3.1, with $\ell_+ \le j_0$. For $j = 0, \dots, j_0 - 1$, it consists of a fixed boundary formula Q_-, also satisfying Ass. 3.1, which to be applicable will have to be one-sided in the sense of having $\ell_- = 0$. Boundary conditions of this kind, namely identical at all points $0, \ldots, j_0 - 1$, are called **translatory boundary conditions** by Goldberg and Tadmor [Go78,Go81]. Now for GKS-stability, since Q_- applies only at a fixed set of points, it is necessary and sufficient that $\tilde{Q}$ admit no steady state solutions that for $j \geq j_0$ consist of rightgoing modes. That is, we can drop the requirement that the solution on the left is leftgoing. However, since $\ell_- = 0$, Q_- admits no rightgoing solutions anyway, so the change is vacuous. Therefore as before, $\tilde{Q}$ is GKS-unstable if and only if it admits some steady state solution that is outgoing on both sides of $j = j_0$, as in Fig. 6.1.

For problems of both of these kinds one main general result appears in the literature: roughly, *total dissipativity ensures stability*. The original theorem in this direction is due to M. Ciment:

Theorem *[Ci72]. Consider the interface problem of the first paragraph above. Let both Q_- and Q_+ be explicit, two-level formulas consistent with the equation $u_t = au_x$. If Q_- and Q_+ are dissipative (i.e. x-dissipative), $\tilde{Q}$ is GKS-stable.* ∎

A similar result for boundary rather than interface problems was derived a few years later, perhaps independently, by Tadmor and Goldberg [Ta78,Go78,Go81]. We express their results in our terminology, in particular replacing their condition (3.7) with the idea of t-dissipativity (§2.2). For a full statement see Thms. 3.3 and 3.4 of [Go81].

Theorem *[Go81]. Consider the initial boundary value problem of the second paragraph above. Let Q_+ be consistent with $u_t = au_x$ for $a > 0$, and assume that Q_- satisfies the von Neumann condition and a certain solvability condition (Defn. 3.1 of [Go81]), but drop the assumption that it satisfies Ass. 3.1. If Q_- is t-dissipative and either Q_- or Q_+ is x-dissipative, then $\tilde{Q}$ is GKS-stable.* ∎

Obviously these two theorems are related, and by isolating the idea of t-dissipativity, we can bring out the connection and generalize them both. In particular, Ciment's restriction to two-level formulas serves no purpose except to ensure that x-dissipativity will imply t-dissipativity (Thm. 2.2.3). Similarly, Tadmor's assumption is unnecessary that it is Q_- rather than Q_+ that is t-dissipative. We propose the following generalizations. In each of these theorems, Q_- and Q_+ may be explicit or implicit, two-level or multilevel.

Theorem 6.2.1.* *Consider the interface problem described in the first paragraph above. Let Q_- and Q_+ be consistent with $u_t = a_-u_x$ and $u_t = a_+u_x$, respectively, with $a_-a_+ > 0$. If at least one of them is x-dissipative and at least one is t-dissipative, then $\tilde{Q}$ is GKS-stable.*

Theorem 6.2.2.* *Consider the initial boundary value problem described in the second paragraph above. Let Q_+ be consistent with $u_t = au_x$ for $a > 0$. If at least one of Q_- and Q_+ is x-dissipative and at least one is t-dissipative, then $\tilde{Q}$ is GKS-stable.*

Proofs. Consider first the case $\ell_+ = r_- = 1$, which covers interfaces between typical three-point formulas. Given z with $|z| \geq 1$, let κ_- and κ_+ denote the κ values for the unique leftgoing and rightgoing modes admitted by Q_- and Q_+, respectively. The abrupt-change interface imposes the condition

$$\kappa_- = \kappa_+ \tag{6.2.1}$$

for a steady-state solution; call this number κ. Now since the signals are outgoing from the interface, we must have $|\kappa_-| \geq 1 \geq |\kappa_+|$, hence $|\kappa| = 1$, and the von Neumann condition for Q_+ then implies $|z| = 1$ also (Thm. 2.2.1). Since one scheme is x-dissipative, these equalities imply $\kappa = 1$. Since one scheme is t-dissipative, this implies further $z = 1$. Now by the consistency assumption, the only signal with $z = \kappa = 1$ is strictly leftgoing on the right of the interface in the initial boundary value problem, while in the interface problem, either it is strictly leftgoing there (case $a_-, a_+ > 0$), or .. is strictly rightgoing but so is the solution on the left of the interface (case $a_-, a_+ < 0$). In any case there can be no unstable solution of the kind illustrated in Fig. 6.1.

Now consider the general problem, in which Q_- and Q_+ have arbitrary stencil parameters ℓ_-, r_- and ℓ_+, r_+. In §3.2 we examined a general abrupt-change interface of this kind in the context of reflection coefficients. Given z with $|z| \geq 1$, let $\kappa_1^-, \ldots, \kappa_{r_-}^-$ denote the leftgoing κ values for Q_-, and $\kappa_1^+, \ldots, \kappa_{\ell_+}^+$ the rightgoing κ values for Q_+. Eq. (3.2.16) showed that an outgoing solution as in Fig. 6.1 exists if and only if the equation

$$VA = 0 \tag{6.2.2}$$

has a solution $A \neq 0$, where V is the van der Monde matrix of size $\ell_+ + r_-$ formed from $\{\kappa_i^-\} \bigcup \{\kappa_i^+\}$, and A is a vector of the same length. We assumed in §3.2 that

*See the qualification in the final paragraph of the proofs.

the sets $\{\kappa_j^-\}$ and $\{\kappa_J^+\}$ had all distinct elements, but if several κ's coalesce, (6.2.2) remains valid if V becomes a confluent Vandermonde matrix involving entries $\kappa^l j^k$. Now a Vandermonde matrix (confluent or not) is singular if and only if it has a repeated column. Therefore (6.2.2) has a nonzero solution if and only if $\kappa_j^- = \kappa_J^+$ for some j and J. From here the proof proceeds as before, beginning with (6.2.1).

These arguments complete the proof, except for one qualification. In the GKS theory that we are appealing to (Thm. 4.3.1), the assumption was made that the difference model is either x-dissipative or strictly nondissipative. This restriction, if it cannot be dispensed with, reduces our theorems to the situation in which both Q_- and Q_+ are x-dissipative. However, we believe that the arguments of [Gu72] can be modified to cover the case in question (see the footnote to Thm. 4.3.1). ∎

Example 6.1. Let Q be a model of $u_t = u_x$ on $(-\infty, \infty)$ consisting of $Q_+ =$ LLd for $j \geq j_0$ coupled with $Q_- =$ BE for $j < j_0$. In §2.2 we have seen that BE is t-dissipative and LLd is x-dissipative. By Thm. 6.2.1, Q is therefore GKS-stable. Coult's theorem does not cover this case.

Example 6.2. Let Q be a model of $u_t = u_x$ on $[0, \infty)$ consisting of $Q_+ =$ LxF (Lax-Friedrichs; see App. A) for $j \geq j_0 \geq 1$ together with the boundary formula

$$v_j^{n+1} = \frac{1}{2}(v_j^{n-1} + v_{j+1}^{n-1}) \tag{6.2.3}$$

for $0 \leq j < j_0$. Since LxF is a two-level formula, it is t-dissipative by Thm. 2.2.3 and it is easy to see that (6.2.3) is x-dissipative. Therefore $\tilde{Q}$ is GKS-stable. (Compare Thm. 4.5.2.) The Goldberg and Tadmor theorem does not cover this case.

6.3 Two interfaces: dissipativity is not enough

The theorems of the last section are appealingly simple, but limited to models containing only one interface. In practice this covers most first- and second-order schemes, which usually have $\ell, r \leq 1$, but very few higher-order ones. It has been thought by some researchers that for multi-interface problems too, dissipativity must ensure stability, and in fact this claim is stated as Thm. 3.1 in [Ol79]. However, dissipativity is in fact *not* enough, and this becomes quite obvious when one thinks in terms of rightgoing and leftgoing solutions and reflection coefficients. In this section we will first explain schematically how instability can come about in multi-interface problems, and then prove it by means of a counterexample.

Here is the explanation. Consider a steady state solution at a GKS-stable interface,

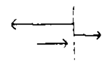


FIG 6.2

where the lengths of the arrows are related somehow to the amplitudes or energy fluxes of the corresponding signals. Now from §§3–5, we know that GKS-stability does not imply that the interface conserves energy, or amplitude, or anything else. It implies that the reflection coefficients are finite, but not that they have moduli less than or equal to 1. Thus as suggested in the figure, a *small* incident wave *may cause* a large reflected wave; it is only a configuration exhibiting reflected energy in the absence of *any* incident energy that would constitute instability.

Now suppose $\tilde{Q}$ contains two stable but nonconserving interfaces of the above sort separated by a fixed number of grid points Δj. Then it may happen that each one stimulates the other's reflected and transmitted energy.

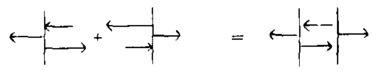


FIG 6.3

If the two-interface system, including all the grid points in between, is thought of as a *single more complicated interface*, as suggested in Fig. 6.3, then the configuration shown generates outgoing waves without any stimulation by incoming ones. Therefore it is GKS-unstable.

Example 6.3: an unstable combination of dissipative stable formulas

This example is extremely contrived, but it yields a strong result. The scheme $\tilde{Q}$ will be an initial boundary value problem model consisting of three dissipative difference formulas Q_0, Q_1, and Q_2, which are applied at $j = 0$, $j = 1$, and $j \geq 2$, respectively. Thus in this example the two interfaces Q_0–Q_1 and Q_1–Q_2 are only one grid step apart. Each formula Q_k is consistent with $u_t = u_x$, and Q_0 and Q_2 are t-dissipative as well as x-dissipative. Yet the combination $\tilde{Q}$ is strongly unstable in that it admits an exponentially growing eigensolution of Godunov-Ryabenkii type, i.e. with $|z| > 1$.

We start from an intended normal mode and build the difference schemes in such as way as to make it indeed an eigensolution of $\hat{Q}$. We will take

$$\lambda = \frac{1}{8}, \qquad z = \frac{129}{128},$$

and aim for the normal mode shown in Fig. 6.4:

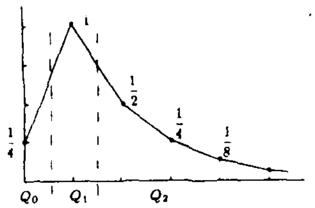


FIG. 6.4

INTERIOR FORMULA: UPWIND DIFFERENCE PLUS DISSIPATION. Q_2 is defined by

$$v_j^{n+1} = v_j^n + \lambda(v_{j+1}^n - v_j^n) + \frac{9\lambda}{8}(v_{j+1}^n - 2v_j^n + v_{j-1}^n) \qquad j \geq 2.$$

With $\lambda = \frac{1}{8}$ this has the characteristic equation

$$z = 1 + \frac{1}{8}(\kappa - 1) + \frac{9}{64}(\kappa - 2 + \frac{1}{\kappa})$$

$$= \frac{38}{64} + \frac{17\kappa}{64} + \frac{9}{64\kappa}.$$

From this formula one may readily verify that $|\kappa| = 1$ implies $|z| \leq 1$, with equality only for $\kappa = z = 1$. This shows that Q_2 is Cauchy stable and totally dissipative. The formula also confirms that for $\kappa = \frac{1}{2}$, as in the mode we have chosen, $z = 129/128$.

LEFTMOST FORMULA: COMBINATION OF UPWIND DIFFERENCES. Q_0 is defined by

$$v_0^{n+1} = v_0^n + \frac{\lambda}{8}\left(\frac{v_2^n - v_0^n}{2}\right) + \frac{7\lambda}{8}\left(\frac{v_3^n - v_0^n}{3}\right).$$

With $\lambda = \frac{1}{8}$ this has the characteristic equation

$$z = 1 + \frac{1}{128}(\kappa^2 - 1) + \frac{7}{192}(\kappa^3 - 1)$$

$$= \frac{367}{384} + \frac{3\kappa^2}{384} + \frac{14\kappa^4}{384}.$$

This formula implies that Q_0, like Q_2, is Cauchy stable and totally dissipative, since $|z| < 1$ for $|\kappa| = 1$ except when $\kappa^2 = \kappa^3 = 1$, hence $\kappa = 1$. Applying it to the chosen normal mode gives again the growth factor $z = 129/128$.

MIDDLE FORMULA: LEAP FROG PLUS IMPLICIT DISSIPATION. Q_1 is defined by

$$v_1^{n+1} = v_1^{n-1} + \lambda(v_2^n - v_0^n) + \epsilon(v_2^{n+1} - 2v_1^{n+1} + v_0^{n+1})$$

with $\epsilon > 0$ (cf. LFd), which for $\lambda = \frac{1}{8}$ has the characteristic equation

$$z(1 - \epsilon(\kappa - 2 + 1/\kappa)) = \frac{1}{z} + \frac{1}{8}(\kappa - 1/\kappa).$$

For $|\kappa| = 1$ and $\kappa \neq 1$ this becomes

$$Mz - \frac{1}{z} = \frac{1}{8}(\kappa - \frac{1}{\kappa})$$

with $M > 1$, and as the right hand side is pure imaginary it can equal the left hand side only when $|z| < 1$, so the scheme is Cauchy stable and dissipative. (The possibility $z = \pm i$ must be disposed of separately.) We are entirely done if $\epsilon > 0$ can be chosen so that when the characteristic equation is applied to the normal mode of Fig. 6.4, the growth will be $z = 129/128$. For this one needs

$$z = \frac{1}{z} + \frac{1}{8}\left(\frac{1}{4}\right) + \epsilon z\left(\frac{-5}{4}\right),$$

that is,

$$\epsilon = \frac{z - 1/z - 1/32}{-5z/4} = \frac{1036}{83205} \approx .01245117.$$

According to these definitions Q_1 and Q_2 each have one leftgoing and rightgoing mode for all $|z| \geq 1$, while Q_0 has three leftgoing modes. All together, therefore, the proposed normal mode has the schematic form

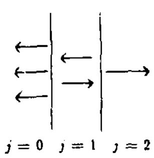


FIG. 6.5

This is consistent with the general form of a GKS instability illustrated in Fig. 6.1.

DEMONSTRATION 6.1. To confirm the above analysis, the model $\hat{Q}$ was applied on a grid $h = 1/100$, $k = 1/800$ on $[0,2]$ with initial data

$$v_j^0 = 1 - \frac{jh}{2},$$

with inflow boundary condition $v_{200}^{n+1} = 0$. The following unstable growth was observed:

t	n	$\|v^n\|_{RMS}$	Ratio
0	0	.5781	
1	800	.2010	.348
2	1600	8.404	41.8
3	2400	4.251×10^3	506
4	3200	2.149×10^6	506

TABLE 6.1

The ratio rapidly approaches the predicted value $(129/128)^{800} \approx 505.6$. A plot of the computed distribution also shows exactly the form of the predicted normal mode.

6.4 Two interfaces: stability and reflection coefficients

The example of the last section showed that when two or more GKS-stable interfaces *interact*, the presence of reflection coefficients greater than 1 in modulus may cause the combination to be unstable. Here we will show, conversely, that if the moduli are not greater than 1, this implies stability. The problem we apply this idea to comes from a paper of Beam, Warming, and Yee [Be81], in which they motivate and define the notion of *P-stability*. (Beam et al. do not argue by means of reflection coefficients.) We will reproduce and extend their main results.

The background to [Be81] is as follows. In studying certain fluid flow problems numerically on an interval $[0,1]$, Beam et al. applied time-dependent finite-difference models for the purpose of determining steady-state solutions, i.e. $t \to \infty$. Since they had relatively little interest in modeling the transient behavior accurately, it was natural to consider large time steps, hence large mesh ratios λ, and even to consider the limit $\lambda \to \infty$. When they did this, they observed that in some cases, large values of λ led to models that admitted exponentially growing solutions, even though each

boundary individually was GKS-stable. Such exponential growth does not violate GKS-stability, *provided it does not become more severe as the mesh is refined*, but it is fatal for computations in which one wants a meaningful limit as $t \to \infty$.

Therefore Beam et al. defined

Defn. [Be81]. A difference model is **P-stable** if it is GKS-stable, and furthermore, for all $h > 0$ it admits no eigensolutions with $|z| > 1$. //

By an eigensolution, we mean here an eigensolution in the standard functional analytic sense of the operator representing the entire difference scheme, including boundary conditions at both ends. This definition rules out the troublesome exponential growth. *The trouble is* that it is not a stability definition of the usual sort, since it is not connected with any estimate like (4.2.5) or (4.3.1). However, the complexity of the GKS theory in general, and of (4.3.1) in particular, *suggest* that it may sometimes be useful to discusss practical stability criteria without waiting for a complete theory to justify them. (This is what we did in §5.) In experiments on their original nonlinear fluids problem, Beam et al. found that P-stability is a reliable guide to observed success of the computation.

The observation that GKS-stable strip models may admit exponentially growing solutions is not new, and in fact §7 of [Gu72] is devoted to this phenomenon. For a particular example involving a 2×2 system, that section derives conditions in terms of λ and the number of grid points between the boundaries for there to be no growing eigensolutions. The contribution of [Be81] is that it applies *similar* ideas in a more realistic context, and in particular it derives P-stability results for an interesting class of models based on A-stable formulas.

We will now derive P-stability results by means of reflection coefficients. Consider the two-interface geometry shown in Fig. 6.6, on which a composite difference model $\hat{Q}$ is applied.

$\cdots \times \quad \times \quad \big| \quad \times \quad \times \quad \times \quad \times \quad \big| \quad \times \quad \times \cdots$

$j = 0 \qquad\qquad j = J$

FIG 6.6

At points $j = 1, \ldots, J-1$, with $J \geq 1$, $\hat{Q}$ consists of a constant-coefficient, Cauchy-stable difference formula Q_0. For simplicity we assume that Q_0 is a three-point scalar formula satisfying Ass. 3.1 with $\ell = r = 1$. (The ideas to follow can all be extended to more complicated difference models, including systems as well as scalars.) For $j \leq 0$ and $j \geq J$, two additional Cauchy stable formlas Q_- and Q_+ are applied. Though the figure illustrates the pure interface case $-\infty < j < \infty$, we will permit one or both interfaces to degenerate to boundaries, as in §6.2—in which case Q_- or Q_+ becomes one-sided, and we cease to require Cauchy stability for that formula. If both interfaces are boundaries, we speak of the "boundary case"; if at least one is an internal interface, we speak of the "interface case".

Suppose that $\tilde{Q}$ admits a steady-state solution with $|z| \geq 1$. For $0 \leq j \leq J$ it will necessarily have the form

$$v_j^n = z^n(\alpha\kappa_\ell^j + \beta\kappa_r^j) \qquad (0 \leq j \leq J). \tag{6.4.1}$$

Let A_1 and A_2 (functions of z) denote the reflection coefficients at the left and right, respectively, as considered in §3. That is, A_1 denotes the ratio of amplitudes of the rightgoing signal to the leftgoing one at $j = 0$, and analogously for A_2. Then (6.4.1) implies that α and β satisfy

$$\beta = A_1\alpha, \qquad \alpha\kappa_\ell^J = A_2\beta\kappa_r^J. \tag{6.4.2}$$

(We permit the GKS-unstable possibilities $A_1 = \infty$ and $A_2 = \infty$.) If we set $\alpha = 1$, then $\beta = A_1$, and (6.4.1) becomes

$$v_j^n = z^n(\kappa_\ell^j + A_1\kappa_r^j).$$

But the second equation of (6.4.2) implies further

$$A_1A_2(\kappa_r/\kappa_\ell)^J = 1. \tag{6.4.3}$$

We can interpret this as follows: if at a fixed time step we trace the rightgoing mode from $j = 0$ to $j = J$, reflect it by a factor A_2 to a leftgoing mode, trace this back to $j = 0$, and reflect it by A_1 to the rightgoing signal again, then we must have the same value we started with.

In (6.4.3), all of the quantities $A_1, A_2, \kappa_\ell, \kappa_r$ depend on z. This equation contains all the information relevant to stability analysis: $\tilde{Q}$ admits an eigensolution for a given $z \in \mathbb{C}$ if and only if (6.4.3) is satisfied for that z. Determining whether this is so for a range of values of z may be difficult. The advantage of (6.4.3) is that it permits one

to make simpler inferences if the reflection coefficients are well behaved. Here is the most natural such result:

Theorem 6.4.1. *Let the two-interface model $\tilde{Q}$ be defined as above. If $|A_1| \leq 1$ and $|A_2| \leq 1$ for all z with $|z| \geq 1$, then $\tilde{Q}$ admits no eigensolutions with $|z| > 1$. If in addition $|A_1| < 1$ or $|A_2| < 1$ or both for each such z, then $\tilde{Q}$ admits no eigensolutions or generalized eigensolutions with $|z| \geq 1$.*

Proof. Since $|\kappa_r| \leq 1 \leq |\kappa_\ell|$ for $|z| \geq 1$, one has $|(\kappa_r/\kappa_\ell)^J| \leq 1$, and the second statement is an immediate consequence of (6.4.3). For the first, one uses the additional fact that Cauchy stability implies $|\kappa_r| < 1 < |\kappa_\ell|$ for all $|z| > 1$, so that one has $|(\kappa_r/\kappa_\ell)^J| < 1$. ∎

Remark. This result holds even if one or both interfaces are GKS-unstable (cf. Observation 5.4).

Theorem 6.4.1 yields a simple proof of the first main theorem of Beam, et al. Recall the notions of *three-point linear multistep formulas* and *A-stability* described in §2.4.

Theorem 6.4.2 *([Be81], Thm. 4.1). Let $u_t = u_x$ be modeled on $[0,1]$ by a difference scheme $\tilde{Q}$ consisting of a three-point linear multistep formula Q_0 for $j = 1, \ldots, J-1$, together with boundary conditions $v_J^{n+1} = 0$ at $x = 1$ and $(q-1)$st-order space extrapolation S (3.2.29) at $x = 0$ for some $q \leq J$. If Q_0 is A-stable, then $\tilde{Q}$ is P-stable.*

Proof. From (3.2.31) or by a simple computation, the left-hand reflection coefficient is

$$A_1 = -\left(\frac{1-\kappa_\ell}{1+\kappa_\ell}\right)^q \kappa_\ell^q. \tag{6.4.4}$$

By (2.4.14), the A-stability implies $\operatorname{Re}\kappa_\ell \geq 0$, and it follows that the term in parentheses has modulus at most 1. This implies

$$|A_1| \leq |\kappa_\ell|^q \qquad \text{for } |z| \geq 1. \tag{6.4.5}$$

Moreover, the nonvanishing of the denominator of (6.4.4) implies by Thm. 4.3.2 that the boundary at $j = 0$ is GKS-stable.

The right-hand condition $v_J^n = 0$ is trivially GKS-stable. This condition is equivalent to the imposition of a reflection coefficient

$$A_2 = -1. \tag{6.4.6}$$

Since each boundary is GKS-stable, $\tilde{Q}$ is GKS-stable by Thm. 5.2.1. It remains to show that there are no eigensolutions with $|z| > 1$. By the first statement of Thm. 6.4.1 together with eqs. (6.4.5) and (6.4.6), we would be done if the inequality $|\kappa_\ell| \le 1$ were valid. Since $|\kappa_\ell| > 1$ for $|z| > 1$, the situation is not quite this simple, but the idea of Thm. 6.4.1 still applies, and with the use of the fact $J \ge q$, the proof can be finished in either of two ways. Bypassing Thm. 6.4.1, one can return to (6.4.3) and obtain immediately the contradiction

$$1 = |A_1 A_2 (\kappa_r/\kappa_\ell)^J| \le |\kappa_\ell|^q |\kappa_r|^J |\kappa_\ell|^{-J} = |\kappa_r|^{2J-q} < 1$$

for any solution with $|z| > 1$. (For the second equality we have made use of (2.4.8).) Alternatively, one can shift the interface by renumbering the indices so that the old $j = q$ becomes a new $j' = 0$, after which A_1' will satisfy

$$|A_1'| \le |\kappa_\ell|^q |\kappa_r/\kappa_\ell|^q \le 1$$

for $|z| \ge 1$. Then Thm. 6.4.1 applies directly. ∎

Theorem 6.4.2 has the following simple, if not very practical, analog for problems in which three A-stable formulas are separated by abrupt-change interfaces.

Theorem 6.4.3. *Let $u_t = a u_x$ be modeled on $(-\infty, \infty)$ by the two-interface model $\tilde{Q}$ of Fig. 6.6, composed of consistent A-stable three-point linear multistep fomulas Q_-, Q_0, and Q_+. Then $\tilde{Q}$ admits no eigensolutions or generalized eigensolutions with $|z| \ge 1$, except possibly an eigensolution or generalized eigensolution with $|z| = 1$ that is non-strictly leftgoing in $j \le 0$ and non-strictly rightgoing in $j \ge J$.*

Proof. To begin with we have a problem with interfaces at $j = 0$ and $j = J \ge 1$, but as in the last proof, let us shift the indices so that the interfaces lie at $j = \frac{1}{2}$ and $j = J - \frac{1}{2}$. This will multiply both reflection coefficients A_1 and A_2 by the factor $\sqrt{\kappa_r/\kappa_\ell}$. Now by (3.2.5), taking into account the shift of indices, A_2 has the value

$$A_2 = -\frac{\kappa_t - \kappa_r}{\kappa_t - \kappa_\ell}. \tag{6.4.7}$$

(Here ℓ, r, and t stand for "leftgoing", "rightgoing", and "transmitted"; these abbreviations differ from those of (3.2.5), where i stands for "incident" and r for "reflected".) Assume without loss of generality $a > 0$. Then by Thm. 2.4.1, κ_r and κ_t lie in the closed left half of the unit disk ($|\kappa| \le 1$, $\operatorname{Re}\kappa \le 0$), while κ_ℓ lies outside the disk in the right half plane ($|\kappa| \ge 1$, $\operatorname{Re}\kappa \ge 0$). The configuration is indicated in Fig. 6.7:

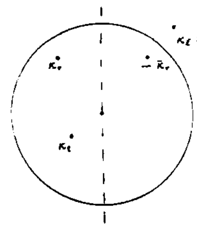


FIG. 6.7

By simple geometry there follow the inequalities

$$|\kappa_t - \kappa_r| \le |\kappa_t - (-\bar{\kappa}_r)| \le |\kappa_t - \kappa_\ell|;$$

the first two terms are equal if and only if $\operatorname{Re}\kappa_r = 0$ or $\operatorname{Re}\kappa_t = 0$, and the latter two if and only if $|\kappa_r| = 1$. Applying these facts to (6.4.7) gives

$$|A_2| \le 1, \tag{6.4.8}$$

with equality if and only if either $\operatorname{Re}\kappa_t = 0$ and $|\kappa_r| = 1$, or $\kappa_r = \pm i$. Obviously one must then have $|A_1| \le 1$ for the reflection coefficient at the left-hand interface, also, with equality under analogous conditions.

By the first statement of Thm. 6.4.1, (6.4.8) and the corresponding bound $|A_1| \le 1$ imply that $\tilde{Q}$ admits no eigensolutions with $|z| > 1$. By the second statement of that theorem, there can be no eigensolutions or generalized eigensolutions for $|z| = 1$ either unless $|A_1| = |A_2| = 1$, which by the remarks above implies either $\operatorname{Re}\kappa_t = 0$ or $\kappa_r = \kappa_\ell = \pm i$, and analogously at the left-hand interface. To complete the proof it is therefore enough to show that each of these last two possibilities implies that the transmitted signal $\kappa_t^j z^n$ is non-strictly rightgoing. In the first case, $\operatorname{Re}\kappa_t = 0$, this is immediate: either $|\kappa_t| < 1$, and the signal is evanescent (position (7) in Table 2.1), or $\kappa_t = \pm i$, and it is a stationary wave with $C = 0$ (position (5)). In the second case, $\kappa_r = \kappa_\ell = \pm i$, then we are done as before if it happens that $\kappa_t = \kappa_r = \pm i$ also. On the other hand if $\kappa_t \ne \kappa_r$, then by (6.4.7), $A_2 = -1$, in which case the leftgoing and rightgoing components cancel each other and there is no generalized eigensolution after all. ∎

Now let us return to the two-boundary problem. The more complicated results of [Be81] involve strongly A-stable schemes used in combination with the boundary formula ST (3.2.32) at $j = 0$. It is in this case that Beam, et al. observed P-instability. For odd values of J, the formulas they considered appeared to be P-stable, but for even values P-stability held only if one restricted attention to values $J \ge f(\lambda)$ for a certain function $f(\lambda)$, increasing somewhat faster than linearly with λ.

We can explain and extend these results by means of reflection coefficients. First we establish GKS-stability for strongly A-stable formulas:

Theorem 6.4.4 *([Be81], Thm. 4.2). Let $\tilde{Q}$ and Q_0 be defined as in Thm. 6.4.2, except with S replaced by the $(q-1)$st-order space-time extrapolation boundary condition ST (3.2.32) at $j = 0$. If Q_0 is strongly A-stable, then $\tilde{Q}$ is GKS-stable.*

Proof. For GKS-stability of $\tilde{Q}$ we need only prove GKS-stability for the interfaces $j = 0$ and $j = J$ independently, by Thm. 5.2.1, and we considered the latter interface already in the proof of Thm. 6.4.2. At $j = 0$, (3.2.34) gives the reflection coefficient

$$A_1 = -\left(\frac{z - \kappa_\ell}{1 + z\kappa_\ell}\right)^q \kappa_\ell^q. \tag{6.4.9}$$

We need to show that the denominator cannot vanish, i.e. $z\kappa_\ell \neq -1$ for all z with $|z| \geq 1$. By Thm. 2.4.1, the strong A-stability implies $|\kappa_\ell| > 1$ for all z with $|z| \geq 1$ except in the case $\kappa_\ell = 1$. By Thm. 2.4.2, Q_0 is ℓ-dissipative, and therefore with $\kappa_\ell = 1$ one has either $z = 1$ or $|z| < 1$. Neither of these possibilities permits $z\kappa_\ell = -1$. ∎

Now let us show that although the model based on ST is GKS-stable, it can no longer be expected to be P-stable, at least when the mesh ratio is large. Assume $\lambda \gg 1$. Then by (2.4.3), one has $\kappa = 1/\kappa + O(1/\lambda)$, hence

$$\kappa_r = -1 + O(\frac{1}{\lambda}), \qquad \kappa_\ell = 1 + O(\frac{1}{\lambda}). \tag{6.4.10}$$

In particular this will hold for $z \approx -1$. But for these values, the denominator of (6.4.9) has magnitude $O(\lambda^{-q})$, which implies that the reflection coefficient will be very large:

$$|A_1| \geq \text{const. } \lambda^q.$$

This explains the observed P-unstable behavior. For large λ, the left boundary of $\tilde{Q}$ is "nearly GKS-unstable" it admits a rightgoing signal $(\kappa_r, z) \approx (-1, -1)$ stimulated by only a very weak leftgoing signal $(\kappa_\ell, z) \approx (1, -1)$.

F[illegible]-stability to be assured by arguments based on (6.4.3), the attenuation $|\kappa_r/\kappa_\ell|$ [illegible] the interior must be strong enough to more than balance the amplification due to A_1. Since $|\kappa_r/\kappa_\ell| = 1 - O(1/\lambda)$, by (6.4.10), this will require

$$(1 - \frac{1}{\lambda})^J \leq \frac{\text{const.}}{\lambda^q},$$

or by taking the logarithms of both sides,

$$J \log(1 - \frac{1}{\lambda}) \leq \text{const. } q \log \frac{1}{\lambda},$$

hence

$$J \geq \text{const. } q\,\lambda \log \lambda. \tag{6.4.11}$$

This kind of relationship between λ and J is just what Beam, et al. observed in practice.

By performing the above estimates carefully, one could derive a precise condition like (6.4.11) that would be sufficient for P-stability. This would complement nicely the third main result of [Be81], Thm. 4.3 there, which gives a bound much like (6.4.11) that is *necessary* for P-stability when J is even.

It remains to give an explanation of the odd-even effect described above. We have shown that for J smaller than the order of magnitude indicated by (6.4.11), the left hand side of (6.4.3) cannot be *guaranteed* to have modulus less than 1, and so an argument balancing reflection and attenuation does not rule out growing eigensolutions. However, from the above results it follows that in the region $z \approx -1$, $\kappa_r \approx -1$, $\kappa_\ell \approx 1$, where A_1 is large, A_1 will be approximately *negative*. That is to say, it will have large negative real part and relatively small imaginary part. Combining this fact with (6.4.6) shows that the left hand side of (6.4.3) has sign approximately

$$(-1)(-1)(-1)^J = (-1)^J.$$

If J is odd, the sign is negative and (6.4.3) cannot hold, despite the large reflection coefficient. This is why P-instability does not occur when J is odd.

6.5 Growth rates for two-interface problems

In this section we continue the pattern of argument of §6.3, in which necessary conditions for instability were derived by balancing amplification by reflection at the interfaces against attenuation in the interior (eq. (6.4.3)). The difference is that here the aim is not primarily to rule out solutions with $|z| > 1$, but to estimate their rates of growth when they do occur.

Both §6.3 and §6.4 were concerned with the fact that the combination of two GKS-stable boundaries or interfaces may be unstable. Section 6.3 considered catastrophic instability in the case of fixed separation Δj, and §6.4 considered P-instability in the case of fixed Δx. A third, related *phenomenon is described* by Kreiss in various papers, and we have discussed this in Chapter 5: if a GKS-*unstable* boundary is used in a strip problem with fixed Δx, the interaction of the two boundaries may convert an instability with $|z| = 1$ to exponential [Kr71, §2; Kr73, §17]. All of these are just three of a variety of effects which can arise that involve "reflection back and forth" between two boundaries. The analysis in this section will consider phenomena of this sort systematically, to see what kind of reflection is really going on, and what degree of unstable growth, if any, to expect in various circumstances.

Again, consider a two-interface constant-coefficient model $\tilde{Q}$ of the kind described in §6.4 and illustrated in Fig. 6.6, with either J constant (the "fixed Δj" case) or $Jh = \Delta x$ constant (the "fixed Δx" case). From (6.4.3), we know that a steady-state solution (6.1.1) for some $z \in \mathbb{C}$ can exist only if

$$|\kappa_r/\kappa_\ell|^{J} = |A_1 A_2| \tag{6.5.1}$$

for that z. It is this equation that asserts that attenuation and amplification must balance. For simplicity let us write

$$\kappa = |\kappa_r/\kappa_\ell| \leq 1, \qquad A = |A_1 A_2|. \tag{6.5.2}$$

Then (6.5.1) can be written

$$A \approx \kappa^{-J}. \tag{6.5.3}$$

We use the symbol "$\approx$", without defining it precisely, because throughout this section we will ignore constant factors. (Of course, in (6.5.3) the two sides are *actually* equal.) The pattern of argument we will use is to show that (6.5.3) can hold only when $|z|$ has a certain size, dependent on J. It then follows that one can observe no unstable growth worse than $|z|^n$, for values of $|z|$ in this range. If $|z| = 1 + \epsilon$ with $\epsilon \ll 1$, the rate of growth becomes

$$E(n) \approx |z|^n = e^{n \log(1+\epsilon)} = (\text{const.})^{n\epsilon}, \tag{6.5.4}$$

where $E(n)$ denotes, say, the ℓ_2 norm $\|v^n\|_2$.

Our aim is to find worst-case rates of growth $E(n)$ for various classes of two-interface problems. The worst-case idea amounts to assuming that (6.5.3) is a sufficient as well as *necessary* condition for a steady-state solution to exist. Of course for particular problems this may not be so (we might have $A_1 A_2(\kappa_r/\kappa_\ell)^J = -1$ instead of

(6.4.3), for example), but across classes of problems it should be valid. The set of problems to be considered is defined by the following parameters, to which we have given labels for convenience:

(Δj) fixed Δj ($= J$),
(Δx) fixed Δx (i.e. $J \to \infty$);

(1) $A \leq 1$ (GKS-stable or unstable),
(2) $1 < A < \infty$ (GKS-stable or unstable),
(3) $A = \infty$ with $|z| = 1$ (GKS-unstable),
(4) $A = \infty$ with $|z| > 1$ (GKS-unstable);

(D) $\kappa \leq \kappa_0 < 1$ for $|z| \geq 1$ (Q_0 totally dissipative),
(ND) $\kappa = 1$ for z with $|z| = 1$ (Q_0 nondissipative).

Except in case (4), we assume that no Godunov-Ryabenkii eigen*solution* with $|z| > 1$ is present. As the list suggests, the arguments will depend on A but not on GKS-stability *per se*. This fact supports Obs. 5.4.

We will ignore exceptional cases, such as those involving defective values of κ or z. First we classify the "best" cases ($E \leq$ const.), then the "worst" ones ($E \approx$ (const.)^n), then various cases in between. The results are summarized at the end in Table 6.2.

Case $A \leq 1$

Suppose $A \leq 1$ (case (1)). In Thm. 6.4.1, we have seen that there can be no steady-state solutions with $|z| > 1$. This rules out an exponential growth no matter what *combination of the remaining* parameters above is in effect. We can interpret this in terms of energy moving back and forth between interfaces as follows: an initial perturbation may persist for all time, reflecting back and forth between interfaces, but it will not grow. If Q_0 is totally dissipative, it should die out.

One kind of growth may still be expected. In the case of an interface (not boundary) problem, with Q_0 nondissipative (ND), a signal of the above sort trapped between the interfaces may radiate wavelike energy into the left- or righthand semi-infinite region. This will cause algebraic growth in E. In the fixed Δx case, the growth will *look* qualitatively like $E \approx 1 + \sqrt{t}$, which is not unstable because of the initial magnitude 1. In the fixed Δj case, it will look like $E \approx Jh + \sqrt{t}$, which is unstable, since $Jh \to 0$ as $h \to 0$. The examples of §6.4 illustrate these circumstances.

In summary, the case $A \leq 1$ should grow at worst as follows. The symbol ($\bullet$) is a wild card indicating any of the choices of the parameter in that position.

(Δj)-(1)-(ND): unstable algebraic growth, $E \approx h + \sqrt{t}$ (interface case only);
(Δx)-(1)-(ND): stable algebraic growth, $E \approx 1 + \sqrt{t}$ (interface case only);
$(\bullet)$-(1)-(D): no growth.

Case $A = \infty$ with $|z| > 1$

At the other extreme, suppose that one or both interfaces admits an eigensolution of the Godunov-Ryabenkii kind (case (4)), as described in §4.2, with an infinite reflection coefficient. Since such an interface alone would exhibit growth like (const.)^n, it is natural to expect the same for the two-interface problem, or worse. In fact there can be nothing worse; this follows from the bounded solvability from one time step to the next of any properly defined difference model (Ass. 3.3). It remains just to *confirm* that a steady state solution with $E \approx$ (const.)^n can indeed occur. This raises the question, how can $A = \infty$ and $\kappa > 0$ be reconciled with (6.5.3)?

The answer, which will reappear throughout this section, is that a steady-state solution with two interfaces will not have z equal to the value z_0 for which $A = \infty$, but to a perturbed value z_0'. Assume first $\kappa = 1$, case (ND). Then the perturbation $z_0 \to z_0'$ must be large enough to bring $A(z)$ down to $O(1)$, which means $z_0' - z_0 \approx O(1)$. Since $|z_0| > 1$, however, this is not inconsistent with $|z_0'| > 1$. Hence an exponentially growing solution (6.4.1) may occur.

If Q_0 is totally dissipative (case (D)), growth of the form $E \approx$ (const.)^n will still typically occur, but the perturbation argument may change. In case (Δj) the attenuation $\kappa^J > 0$ is insignificant compared to the reflection coefficient $A(z_0) = \infty$, so $z_0' - z_0 = O(1)$ again. But in case (Δx), $J = O(1/h)$, and κ^J is not bounded away from 0. Assume that for $z \approx z_0$, A looks something like

$$A \approx |z - z_0|^{-1}. \qquad (6.5.5)$$

Then to satisfy (6.5.3) one must have

$$z_0' - z_0 = O(\kappa^J). \qquad (6.5.6)$$

For any reasonable value of J this implies that z_0' will be extremely close to z_0. In other words, the exponential instability admitted by the two-interface system will look almost exactly like the single-interface instability with $z = z_0$.

Of course the two interfaces might interact fortuitously so as to rule out such a solution, as mentioned already. In case (Δj), $\hat{Q}$ would then actually be stable, even

though composed of one or two strongly unstable interfaces. In case (Δx), it would be unstable nevertheless. The reason is that for large J (small h), a perturbation near one interface might grow catastrophically for a time before feeling the influence of the other interface and finally decaying to 0; as $h \to 0$ the catastrophe becomes worse.

We are however concerned with worst-case growth, for which the summary of models with $A = \infty$ for $|z| > 1$ is very simple:

$(\bullet)$-(4)-$(\bullet)$: unstable exponential growth, $E \approx$ (const.)^n.

Case $1 < A < \infty$ or $A = \infty$, fixed Δj

In the first case indicated, one has two nonconserving but possibly GKS-stable interfaces separated by a fixed number of grid points. The example of §6.3 exhibiting growth $E \approx (\frac{129}{128})^n$, was of this kind, namely (Δj)-(2)-(D). Obviously if Q_0 is nondissipative (ND), or if $A = \infty$ instead of $A < \infty$, growth like (const.)^n should still be possible. There is also no distinction here between the boundary and the interface situations.

There are however qualifications for the cases (Δj)-(3)-(D) and (Δj)-(2)-(D). Let Q_0 be fixed and totally dissipative, and suppose first $A = \infty$. As in the last discussion, we are once again led to the perturbation (6.5.6), which is extremely small except when J is near 0. But this time $|z_0| = 1$, so that (6.5.6) implies that $|z_0'|$, although perhaps larger than 1, may be extremely close to it. Therefore the growth, although exponential, will be slow in this case unless $J \approx 0$. On the other hand suppose $A(z) \le A_{max} < \infty$ for some A_{max}. Then (6.5.3) can only hold for J small enough so that $\kappa^J A_{max} \ge 1$, say $J \le J_0$. For practical examples of this type, such as the interaction of GKS-stable interfaces with $Q_0 =$ LW, J_0 usually seems to be 0 - dissipation almost always produces stability. This is why the example of §6.3 had to be so contrived. Unfortunately, it is generally hard to prove that J_0 is so small, even for particular examples.

In summary,

(Δj)-(3)-(ND): unstable exponential growth, $E \approx$ (const.)^n;
(Δj)-(3)-(D): weak exponential growth, $E \approx$ (const.)^n, const. $- 1 \ll 1$;
(Δj)-(2)-(ND): unstable exponential growth, $E \approx$ (const.)^n;
(Δj)-(2)-(D): $E \approx$ (const.)^n for $J \le J_0$, no growth for $J \ge J_0$.

Case $1 < A < \infty$ or $A = \infty$, fixed Δx

The most interesting set of cases remains: those with fixed Δx and either $1 < A < \infty$ (GKS-stable or unstable) or $A = \infty$ but $|z| = 1$ (GKS-unstable). Depending on which kind of interface is present and whether Q_0 is dissipative, four different rates of growth may be expected.

The case (Δx)-(2)-(D) has already been settled by our discussion of the case (Δy)-(2)-(D). There we argued that for $J \geq J_0$, no growth will occur. In the fixed Δx situation, we are only concerned with the limit $J \to \infty$, and so one should expect no growth here. B. Gustafsson has stated a theorem to this effect in [Gu81]. Similar results for particular examples appear in [Gu72], §7.

Suppose that again $1 < A < \infty$, but Q_0 is nondissipative—case (Δx)-(2)-(ND). Whether or not $\hat{Q}$ is GKS-stable, in general it will be susceptible to exponential growth in t (not n), $E \approx (\text{const.})^t$. There are two ways to see this. One is to think of reflections back and forth as t increases. In the worst case, a signal might bounce repeatedly between the two interfaces, increasing in magnitude by a factor $A > 1$ with each circuit. In the fixed Δx case the travel time will be $O(1)$ between bounces, and so one has a growth rate $(\text{const.})^t$. Alternatively, one can argue again by perturbations $z_0 \to z_0'$. If $|z_0| = \kappa(z_0) = 1$, then typically if $|z_0'| \approx 1+\epsilon$, we will have $\kappa(z_0') \approx 1-\epsilon$. Since $J = O(1/h)$, (6.5.3) becomes the condition

$$(1-\epsilon)^{-1/h} \approx A \approx 1, \qquad (6.5.7)$$

which implies $\epsilon \approx h$. In other words, following (6.5.4), we should observe growth like

$$E(n) \approx (1+h)^n = (\text{const.})^t.$$

Note that growth at this rate, although stable (take $\alpha_0 = \text{const.}$ in (4.2.5) or (4.3.1)), does *not* become weaker as $h \to 0$. This contradicts the impression given by Beam, et al in [Be81], but supports their view that a concept like P-stability may be useful.

Consider now the case of Q_0 nondissipative but $A = \infty$. This is the situation mentioned by Kreiss in which a linear instability may be converted to exponential. The exponential growth is however not of type $(\text{const.})^n$, but of the weaker form $J^t \approx (1/h)^t$. We can see this by the usual perturbation argument. Once again, consider $|z_0'| = 1+\epsilon$ and assume that (6.5.5) holds. Then the condition (6.5.3) is

$$(1-\epsilon)^{-1/h} \approx \frac{1}{\epsilon}. \qquad (6.5.8)$$

For this to be satisfied, ϵ will have magnitude $\epsilon \approx h\log(1/h)$. Therefore by (6.5.4) we will observe growth like

$$E(n) \approx (1 + h\log\frac{1}{h})^{(1/h)} \approx (\text{const.})^{t\log(1/h)} \approx (1/h)^t \approx J^t.$$

Compare [Kr71], §2, or [Kr73], eq. (17.10).

Finally, what will happen if $A = \infty$ but Q_0 is totally dissipative? No matter how large J is, it will still be possible to choose $z_0' - z_0$ to satisfy (6.5.6). However, $z_0' - z_0$ will have to be exceedingly small. Eq. (6.5.4) gives the rate of growth

$$E(n) \approx (1+\kappa^J)^n, \qquad (6.5.9)$$

which is exponential for fixed J but with a constant that decreases rapidly with J. In practice this growth will be completely insignificant, and $\hat{Q}$ will exhibit nothing worse than whatever instability is caused by its individually unstable interface (or interfaces). This conclusion applies in particular to two-boundary problems involving borderlne GKS-instabilities of type $|\kappa| < 1$, as discussed in §5.4. This confirms Observation 5.10.

In summary, for these situations of fixed Δx type we have

(Δx)-(2)-(D): no growth
(Δx)-(2)-(ND): stable growth, $E \approx (\text{const.})^t$
(Δx)-(3)-(D): stable growth, extremely weak
(Δx)-(3)-(ND): unstable growth, $E \approx (J)^t$

* * *

Let us summarize the results of this and the previous section. The details have been complicated, but the main idea is simple. For a growing eigensolution to exist, the amplification by reflection at the boundaries must balance the dissipation in the interior, as indicated by (6.5.3). In a model containing an interface with reflection coefficient $A(z_0) = \infty$, one therefore investigates perturbations $z_0 \to z_0'$ to reduce $A(z_0')$ to the right size. The growth rate is then given by $E(n) \approx |z_0'|^n$.

This analysis does not depend on whether any GKS-unstable interfaces are present, confirming Observations 5.4 and 5.10. In a problem of fixed Δx type, GKS-instability at either interface will make itself felt near that interface in the usual way, but interaction of the two interfaces will not worsen the effect unless the reflection coefficient arguments indicate that it should.

We have not discussed borderline GKS-unstable interfaces of type $C = 0$. It turns out that the effect of such a case is typically to introduce a square root on the right hand side of (6.5.5), with similar changes elsewhere. This may weaken unstable growth rates, but it does not change their behavior qualitatively. This is why Obs.

Table 6.2

	fixed Δx		fixed Δj	
	dissipative	nondiss.	dissipative	nondiss.
$A \leq 1$	no growth	$1 + \sqrt{t}$	no growth	$h + \sqrt{t}$
$1 < A < \infty$	no growth	$(\text{const.})^t$	$(\text{const.})^n$	$(\text{const.})^n$
$A = \infty, \lvert z_0\rvert = 1$	weak stable growth	J^t	$(\text{const.})^n$	$(\text{const.})^n$
$A = \infty, \lvert z_0\rvert > 1$	$(\text{const.})^n$	$(\text{const.})^n$	$(\text{const.})^n$	$(\text{const.})^n$

= P-unstable

= GKS-unstable

(assuming individual interfaces are GKS-stable in first two rows)

5.10 differentiates borderline cases of type $|\kappa| < 1$ from those of type $C = 0$. (On the other hand, with $C = 0$ it usually happens that A is finite also, as we saw in §5.3.)

The growth rates we have obtained in this section are summarized in Table 6.2. Once again we emphasize that these rates listed are typical ones, and may be false for special problems; also that our model has involved just two interfaces, a scalar equation, and constant coefficients.

6.6 Three or more interfaces

In this final section we will make some remarks on stability for problems with *three or more interfaces*. Such configurations come up in the design of composite boundary or interface formulas. They are also important to the analysis of adpative mesh refinement schemes, where one would like to be able to derive bounds on growth rates without requiring the number of grid points between mesh refinement interfaces to approach infinity. One might also think of any model with variable coefficients which we have scarcely mentioned in this dissertation as consisting of a series of distinct difference formulas separated by interfaces between each pair of grid points.

The purpose in viewing any of these problems in terms of interfaces is to obtain results by reflection and transmission arguments that might be difficult to obtain otherwise. A particular area where there is a great need for such results is in the study of totally dissipative formulas. For one-dimensional mesh refinement problems, for example, experience shows that virtually any mesh refinement scheme applied with a totally dissipative formula such as LW will be stable. Yet no general theorems along these lines are known, except for the single-interface results discussed in §6.2. As a general rule, although dissipativity helps guarantee the fact of stability, it seems to make the proof of stability more difficult. For example, establishing stability by the energy method for a difference formula with variable coefficients is usually more difficult when the formula is dissipative.

Consider a model Q in which a finite collection of constant difference formula are joined by interfaces separated by a fixed number of grid points. By the results of Chapter 4, we know precisely how to check for GKS-instability of Q, in principle. In each region between interfaces, determine for each z with $|z| \geq 1$ the set of leftgoing and rightgoing signals. At each interface, compute the reflection and transmission matrices that connect them together. Then Thm. 4.3.1 takes the following form. Q

is GKS-stable if and only if the above conditions permit no solution with $|z| \geq 1$ of the kind suggested in Fig. 6.8:

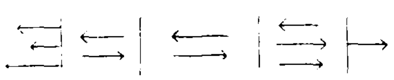


FIG. 6.8

In other words, instability is equivalent to the existence of a solution $v_j^n = z^n \phi_j$ consisting of purely leftgoing components in the left semiinfinite interval, and purely rightgoing components in the right semiinfinite interval, with any combination of left- and rightgoing signals permitted in-between.

In practice, such an analysis will be exceedingly difficult. Therefore one asks, what simple properties of the reflection and transmission coefficients might guarantee that no solution of the form of Fig. 6.8 can occur? As with the models with one or two interfaces considered earlier, an obvious property to look for is conservation of energy. Here is an abstract formulation of the kind of argument that might be used.

Suppose that between each pair of points $(j-1)h, jh$ one can define a *net energy flux* Φ_j (to the right), satisfying the following two properties for all z with $|z| \geq 1$:

(i) $\Phi_j \leq 0$ (resp. ≥ 0) in a region were only leftgoing (resp. rightgoing) modes are present;

(ii) $\Phi_j \geq \Phi_{j+1}$ if jh is not an interface.

For an interface at jh to "conserve energy" with respect to Φ then means simply that *(ii)* holds at that interface as well as at non-interface points. One obtains immediately the following sufficient condition for stability:

Theorem 6.6.1. *Let Φ satisfy conditions (i) and (ii) above. If (ii) holds also at all interface points jh, and in addition at least one inequality in (i) or (ii) is strict, then $\bar{Q}$ is GKS-stable.*

Proof. If there exists an unstable eigensolution or generalized eigensolution, let the corresponding values Φ_j be computed. Let $j_- h$ and $j_+ h$ be points lying to the left and the right all of all interfaces, respectively. Then condition *(i)* implies

$$\Phi_{j_-} \leq 0 \leq \Phi_{j_+},$$

while condition *(ii)* implies

$$\Phi_{j_-} \geq \Phi_{j_-+1} \geq \cdots \geq \Phi_{j_+-1} \geq \Phi_{j_+}. \tag{6.6.1}$$

If any of these inequalities is strict, then these two equations are inconsistent. ∎

Unfortunately, for realistic problems it is not an easy matter to derive a function Φ that is conserved at interfaces. For example, let $\bar{Q}$ consist of a three-point linear multistep formula Q for $u_t = u_x$ applied between an arbitrary finite number of "crude mesh refinement" interfaces of the kind described in Example 3.4 of §3.2. (This is equivalent to applying Q on a uniform mesh with a finite number of coefficient changes, with all coefficients positive.) Let Φ be simply the ℓ_2 energy flux $\Phi_r - \Phi_\ell$, where Φ_r and Φ_ℓ are defined as in §3.3. Then we showed in §3.3 that energy is conserved for $|z| = 1$ when an incident signal is present on one side only, and this result can be readily extended to two-sided incidence. From this one can conclude as in Thm. 6.6.1 that $\bar{Q}$ admits no eigensolutions or generalized eigensolutions for $|z| = 1$ with strictly outgoing signals in one or both semiinfinite region. Yet for $|z| > 1$, the argument breaks down, for it turns out that crude mesh refinement interfaces no longer conserve Φ. Thus Thm. 6.6.1 is not applicable, even though one can show that $\bar{Q}$ is ℓ_2-stable here by standard energy method arguments [Ri67].

This example suggests that the ℓ_2 flux may generally be an unworkable choice for Φ. The same conclusion is suggested by the observation of §3.3 that in most problems, the ℓ_2 energy is not conserved even for $|z| = 1$.

Alternatively, in analogy to the developments of §6.2–§6.5, we could base Φ instead on amplitudes. Consider again a problem with only one leftgoing and one rightgoing solution between each pair of interfaces. One natural choice is

$$\Phi_j = |a_r \kappa_r^j| - |a_\ell \kappa_\ell^j|,$$

but this turns out to be no better than the ℓ_2 definition considered above. However, the possibility

$$\Phi_j = \operatorname{sgn}\left(|a_r \kappa_r^j| - |a_\ell \kappa_\ell^j|\right)$$

shows promise. For this measure of Φ to be conserved at an interface means that the interface admits no solutions in which on each side, the radiated wave has larger amplitude than the incident one. This is a natural extension of the arguments involving $|A| \leq 1$ of the last two sections, and there are indications that it may make it possible to prove stability for some realistic problems.

Unfortunately, we have so far obtained no new results with reflection and transmission arguments of this kind. But we believe they are worth investigating, if only because so little is known at present about multi-interface problems.

APPENDIX A. PROPERTIES OF STANDARD DIFFERENCE MODELS

In the following three pages the properties of eleven common scalar difference formulas, mostly models of $u_t = au_x$, are listed. Each entry gives: (1) name, (2) formula, (3) dispersion relation in κ and z, (4) dispersion relation in ξ and ω, (5) group velocity $C(\xi,\omega)$, (6) initial terms of Taylor series $\omega = \omega(\xi)$ for branch through origin ($\approx$ modified equation), (7) orders of dispersion (α) and dissipation (β), (8) x-reversing and/or t-reversing, and (9) x-dissipative and/or t-dissipative. A dispersion plot for the case $a = -1$, $\lambda = .5$ is also shown for the each formula, following the style of Fig. 1.1. Dashed line segments in these plots indicate solutions (ξ,ω) that are real only at isolated points (for formulas with some dissipation).

See the index for pointers to where the properties listed above, and most of these difference formulas, are discussed in the text.

LF — Leap Frog

$v_j^{n+1} - v_j^{n-1} = \lambda a(v_{j+1}^n - v_{j-1}^n)$ $\qquad z - \frac{1}{z} = \lambda a(\kappa - \frac{1}{\kappa})$

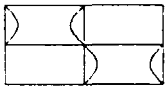
$\sin \omega k = -\lambda a \sin \xi h$ $\qquad C = -a\frac{\cos \xi h}{\cos \omega k}$

$\omega = -a\xi[1 - \frac{1-(\lambda a)^2}{6}(\xi h)^2 + \frac{1-10(\lambda a)^2+9(\lambda a)^4}{120}(\xi h)^4 + \cdots]$

$\alpha = 3$ $\qquad \beta = \infty$

x-reversing $\qquad$ t-reversing

not x-dissipative $\qquad$ not t-dissipative



CN — Crank-Nicolson

$v_j^{n+1} - v_j^n = \frac{\lambda a}{4}(v_{j+1}^{n+1} - v_{j-1}^{n+1} + v_{j+1}^n - v_{j-1}^n)$ $\qquad \frac{z-1}{z+1} = \frac{\lambda a}{4}(\kappa - \frac{1}{\kappa})$

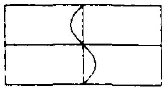
$2\tan\frac{\omega k}{2} = -\lambda a \sin \xi h$ $\qquad C = -a\cos\xi h \cos^2\frac{\omega k}{2}$

$\omega = -a\xi[1 - \frac{1+(\lambda a)^2/2}{6}(\xi h)^2 + \frac{2+10(\lambda a)^2+3(\lambda a)^4}{240}(\xi h)^4 + \cdots]$

$\alpha = 3$ $\qquad \beta = \infty$

x-reversing $\qquad$ not t-reversing

not x-dissipative $\qquad$ t-dissipative



LF4 — Fourth-order Leap Frog

$v_j^{n+1} - v_j^{n-1} = \lambda a[\frac{4}{3}(v_{j+1}^n - v_{j-1}^n) - \frac{1}{6}(v_{n+2}^n - v_{n-2}^n)]$ $\qquad z - \frac{1}{z} = \frac{4\lambda a}{3}(\kappa - \frac{1}{\kappa}) - \frac{\lambda a}{6}(\kappa^2 - \kappa^{-2})$

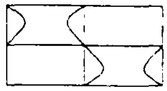
$\sin \omega k = -\frac{4\lambda a}{3}\sin\xi h + \frac{\lambda a}{6}\sin 2\xi h$ $\qquad C = -a\frac{\frac{4}{3}\cos\xi h - \frac{1}{3}\cos 2\xi h}{\cos\omega k}$

$\omega = -a\xi[1 + \frac{(\lambda a)^2}{6}(\xi h)^2 - \frac{4 \quad 9(\lambda a)^4}{120}(\xi h)^4 + \cdots]$

$\alpha = 3$ $\qquad \beta = \infty$

x-reversing $\qquad$ t-reversing

not x-dissipative $\qquad$ not t-dissipative



LW — Lax-Wendroff

$v_j^{n+1} - v_j^n = \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{(\lambda a)^2}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n)$ $\qquad z - 1 = \frac{\lambda a}{2}(\kappa - \frac{1}{\kappa}) + \frac{(\lambda a)^2}{2}(\kappa - 2 + \frac{1}{\kappa})$

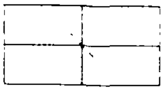
$-i(e^{i\omega k} - 1) = -\lambda a \sin\xi h + 2i(\lambda a)^2\sin^2\frac{\xi h}{2}$ $\qquad C(0,0) = -a$

$\omega = -a\xi[1 - \frac{1-(\lambda a)^2}{6}(\xi h)^2 - i\frac{(\lambda a)-(\lambda a)^3}{8}(\xi h)^3 + \cdots]$

$\alpha = 3$ $\qquad \beta = 4$

not x-reversing $\qquad$ not t-reversing

x-dissipative $\qquad$ t-dissipative



BE — Backwards Euler

$v_j^{n+1} - v_j^n = \frac{\lambda a}{2}(v_{j+1}^{n+1} - v_{j-1}^{n+1})$ — $1 - \frac{1}{z} = \frac{\lambda a}{2}(\kappa - \frac{1}{\kappa})$

$-i(1 - e^{-i\omega k}) = -\lambda a \sin \xi h$ — $C(0,0) = -a$, $C(\frac{\pi}{h}, 0) = a$

$\omega = -a\xi[1 + \frac{i(\lambda a)}{2}(\xi h) - \frac{1+2(\lambda a)^2}{6}(\xi h)^2 + \cdots]$

$\alpha = 3$, $\beta = 2$
x-reversing; not t-reversing
not x-dissipative; t-dissipative

UW — Upwind

$v_j^{n+1} - v_j^n = \lambda a(v_{j+1}^n - v_j^n)$ (if $a > 0$) — $z - 1 = \lambda a(\kappa - 1)$

$e^{i\omega k} - 1 = \lambda a(e^{i\xi h} - 1)$ — $C(0,0) = -a$

$\omega = -a\xi[1 - \frac{i(1-(\lambda a))}{2}(\xi h) - \frac{1-3(\lambda a)+2(\lambda a)^2}{6}(\xi h)^2 + \cdots]$

$\alpha = 3$, $\beta = 2$
not x-reversing; not t-reversing
x-dissipative; t-dissipative

LF^2 — Leap Frog for $u_{tt} = a^2 u_{xx}$

$v_j^{n+1} - 2v_j^n + v_j^{n-1} = (\lambda a)^2(v_{j+1}^n - 2v_j^n + v_{j-1}^n)$ — $(z^{\frac{1}{2}} - z^{-\frac{1}{2}})^2 = (\lambda a)^2(\kappa^{\frac{1}{2}} - \kappa^{-\frac{1}{2}})^2$

$\sin \frac{\omega k}{2} = \pm\lambda a \sin \frac{\xi h}{2}$ — $C = \pm\lambda a \frac{\sin \xi h}{\sin \omega k}$

$\omega = \pm a\xi[1 - \frac{1-(\lambda a)^2}{24}(\xi h)^2 + \cdots]$

$\alpha = 3$, $\beta = \infty$
x-reversing; t-reversing
not x-dissipative; t-dissipative

BX — Box

$v_j^{n+1} + v_{j+1}^{n+1} - v_j^n - v_{j+1}^n = \lambda a(v_{j+1}^n + v_{j+1}^{n+1} - v_j^n - v_j^{n+1})$ — $\frac{z-1}{z+1} = \lambda a(\frac{\kappa-1}{\kappa+1})$

$\tan \frac{\omega k}{2} = -\lambda a \tan \frac{\xi h}{2}$ — $C = -\lambda a \cos^2 \frac{\omega k}{2} / \cos^2 \frac{\xi h}{2}$

$\omega = -a\xi[1 + \frac{1-(\lambda a)^2}{12}(\xi h)^2 + \frac{2-5(\lambda a)^2+3(\lambda a)^4}{240}(\xi h)^4 \cdots$

$\alpha = 3$, $\beta = \infty$
not x-reversing; not t-reversing
not x-dissipative; t-dissipative

MOL — Method of Lines

$\frac{dv_j}{dt} = \frac{a}{2h}(v_{j+1} - v_{j-1})$

$\omega = -\frac{a}{h}\sin \xi h$ — $C = -a \cos \xi h$

$\omega = -a\xi[1 - \frac{1}{6}(\xi h)^2 + \frac{1}{120}(\xi h)^4 + \cdots]$

$\alpha = 3$, $\beta = \infty$
x-reversing; not t-reversing
not x-dissipative; t-dissipative

LFd — Leap Frog with Dissipation

$v_j^{n+1} - v_j^{n-1} = \lambda a(v_{j+1}^n - v_{j-1}^n) - \frac{\epsilon}{16}(v_{j+2}^{n-1} - 4v_{j+1}^{n-1} + 6v_j^{n-1} - 4v_{j-1}^{n-1} + v_{j-2}^{n-1})$

$z - \frac{1}{z} = \lambda a(\kappa - \frac{1}{\kappa}) + \frac{\epsilon}{16z}(\kappa^{\frac{1}{2}} - \kappa^{-\frac{1}{2}})^4$

$C(0,0) = -a$, $C(0, \frac{\pi}{k}) = a$

$\omega = -a\xi[1 + \cdots]$

$\alpha = 3$, $\beta = 4$
not x-reversing; t-reversing
x-dissipative; not t-dissipative

LxF — Lax-Friedrichs

$v_j^{n+1} = \frac{1}{2}(v_{j+1}^n + v_{j-1}^n) + \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n)$ — $z = \frac{1}{2}(\kappa + \frac{1}{\kappa}) + \frac{\lambda a}{2}(\kappa - \frac{1}{\kappa})$

$e^{i\omega k} = \cos \xi h + i\lambda a \sin \xi h$ — $C(0,0) = -a$, $C(\frac{\pi}{h}, \frac{\pi}{k}) = -a$

$\omega = -a\xi[1 - i\frac{1-(\lambda a)^2}{2\lambda a}(\xi h) + \frac{1-(\lambda a)^2}{3}(\xi h)^2 + \cdots]$

$\alpha = 3$, $\beta = 2$
not x-reversing; not t-reversing
not x-dissipative; t-dissipative

APPENDIX B. PROOFS FOR ℓ_2 INSTABILITIES

The purpose of this appendix is to prove Thms. 4.2.3 and 4.2.4. Recall that we are considering a difference model $\tilde{Q}$ for an initial boundary value-problem on $x \geq 0$, $t \geq 0$, and that Q denotes the homogeneous formula applied away from the boundary. The symbols S and $S_{bc}^{(n)}$ denote the operators

$$S: \{v^n, \ldots, v^{n-s}\} \mapsto \{v^{n+1}, \ldots, v^{n-s+1}\} \qquad \text{(homog. bndry data)}, \qquad (B.1)$$

$$S_{bc}^{(n)}: \quad g \mapsto v^n \qquad \text{(homog. initial data)}, \qquad (B.2)$$

with norms induced by ℓ_2 norms on v and g with respect to x and t, respectively (see (4.2.5), (4.2.10), (4.3.2)). We suppose that $\tilde{Q}$ has a strictly rightgoing generalized eigensolution. Then the claims of §4.2 took the form

Thm. 4.2.3. $\|S^n\|_2 \geq \text{const.}\ \sqrt{n}$.

Conjecture. $\|S^n\|_2 \geq \text{const.}\ n$ if an infinite reflection coeff. is present,

Thm. 4.2.4. $\|S_{bc}^{(n)}\|_2 \geq \text{const.}\ n$.

We will prove Thm. 4.2.4 first, then Thm. 4.2.3. The outline of the proofs is as follows. Let h, k be fixed, and consider the *Cauchy problem modeled by Q*, i.e., ignore the boundary to begin with. Construct a wave packet consisting of energy in one or more strictly rightgoing wavelike modes. Initially the wave packet is designed to have width N and lie to the left of $x = 0$. But after a time $O(N)$, the energy will have traveled into $x > 0$, as illustrated in Fig. B.1:

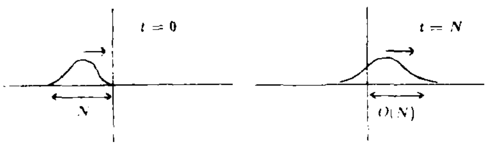


FIG. B.1

In fact if $\|\cdot\|_2$ is the ℓ_2 norm (3.6.2) on $j \geq 0$, we will have $\|v^N\|_2 = O(\sqrt{N})$. Now the key idea is this: *the solution $\{v_j^n\}$ obtained under Q is identical in $j \geq 0$ to the solution $\{\tilde{v}_j^n\}$ obtained if $\tilde{Q}$ is applied with initial data zero and boundary data equal to the numbers $\{g^n\}$ produced when $\{v_j^n\}$ is inserted in the boundary conditions (4.2.3).*

In other words, we can study the initial boundary value problem by means of the Cauchy problem. The distribution $\{v_j^n\}$ would be an exact solution of $\tilde{Q}$ with $g \equiv 0$, as well as of Q, if it happened to satisfy the homogeneous boundary conditions. It doesn't, but it does satisfy $\tilde{Q}$ if we take just the right inhomogeneous "equivalent boundary data." Note the similarity with the arguments of §3.5, and of Fig. B.1 with Fig. 3.9.

If the initial wave packet is made up of a strictly rightgoing generalized eigensolution times a slowly varying envelope, then the homogeneous boundary conditions are nearly satisfied and g is small. In fact we will show that one gets $\|g\|_2 = O(1/\sqrt{N})$, and since $\|v^N\|_2 = O(\sqrt{N})$, Thm. 4.2.4 follows. To prove Thm. 4.2.3, we view the data g not as boundary data but as forcing data F applied at every step, which we then relate to initial data f by the discrete form of Duhamel's principle. In this context g turns out to have norm $O(1)$, so the growth rate is $O(\sqrt{N})$ rather than $O(N)$.

The following argument is divided into four parts. First we consider two-level scalar schemes only, and show in Lemmas B.1 and B.2 that a smooth wave packet behaves as claimed. Then we prove the two theorems for the two-level case. Finally we indicate the extension to general multilevel difference models.

Lemmas on propagation of a smooth wave packet

The core of the proofs is the following lemma, which states that if a wave packet is smooth, then it propagates approximately at the group velocity. Proving this requires the estimation of a Fourier integral that has a standard form: if one divides through by the carrier oscillation $e^{i(\omega t + \xi x)}$, then what remains is the same kind of integral that governs the propagation of a smooth [illegible] difference model of the equation $u_t = -Cu_x$, where [illegible]. A variety of estimates for such integrals are available [illegible]

[illegible] schemes in the maximum norm [illegible]

[illegible]

AD-A119 418 STANFORD UNIV CA DEPT OF COMPUTER SCIENCE F/G 12/1
WAVE PROPAGATION AND STABILITY FOR FINITE DIFFERENCE SCHEMES.(U)
MAY 82 L N TREFETHEN N00014-75-C-1132
UNCLASSIFIED STAN-CS-82-905 NL

2 of 2

END
DATE FILMED
10 82
DTIC

and Wahlbin [Br75], very precise statements can be made about how small the error $v^n(x) - v^0(x + Ct)$ will be, and how this depends on the smoothness of the initial packet and the behavior of the dispersion relation at (ξ_0, ω_0).

However, all we need is a very special case. Therefore rather than appeal to existing theorems, which would introduce undetermined constants and obscure the essential simplicity of what is going on, we give the following argument from first principles.

Let h and k be fixed and let Q be a two-level constant-coefficient Cauchy stable difference formula that admits a solution $e^{i(\omega_0 t - \xi_0 x)}$ with $\xi_0, \omega_0 \in \mathbb{R}$. By Thm. 2.3.1, there exists some group velocity $C \in \mathbb{R}$ such that the dispersion function $\omega = \omega(\xi)$ satisfies

$$\left.\begin{aligned} &\omega = \omega_0 + C(\xi - \xi_0) + \tau(\xi), \\ &|\tau(\xi)| \le M(\xi - \xi_0)^2 \end{aligned}\right\} \quad \forall\, \xi \in \mathbb{R} \tag{B.3}$$

for some constant M. By Cauchy stability, we have $\operatorname{Im} \omega \ge 0$ for all ξ, which implies $\operatorname{Im} \tau(\xi) \ge 0$ also. Since $\eta \mapsto e^{i\eta}$ is a contraction map for $\operatorname{Im} \eta \ge 0$, this implies

$$|e^{i\omega t} - e^{i(\omega_0 + C(\xi - \xi_0))t}| \le t|\tau(\xi)| \le Mt(\xi - \xi_0)^2 \tag{B.4}$$

for any $t \ge 0$.

In what follows the Fourier transform and its inverse are defined by*

$$\hat{f}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x} f(x)\, dx, \qquad f(x) = \int_{-\infty}^{\infty} e^{-i\xi x}\hat{f}(\xi)\, d\xi. \tag{B.5}$$

Lemma B.1. *Let $p(x)$ belong to C_0^2 (twice continuously differentiable with compact support) and satisfy $\hat{p''} \in L_1$. Let Q be applied with initial data*

$$v^0(x) = e^{-i\xi_0 x} p(x).$$

Then for any $n \ge 0$ and any $x \in \mathbb{R}$, $v^n(x)$ satisfies

$$|v^n(x) - e^{i(\omega_0 t - \xi_0 x)} p(x - Ct)| \le Mt\|\hat{p''}\|_1, \tag{B.6}$$

where $t = nk$ and M is the constant of (B.3).

*See the footnote on p. 13 regarding this choice of signs in the exponents.

Proof. Obviously $p \in L_2$, hence $v^0 \in L_2$ also, and we can use Fourier transforms. We get

$$\begin{aligned} v^n(x) &= \int_{-\infty}^{\infty} e^{i(\omega t - \xi x)} \hat{v^0}\, d\xi \\ &= \int_{-\infty}^{\infty} e^{i(\omega t - \xi x)} \hat{p}(\xi - \xi_0)\, d\xi \\ &= \int_{-\infty}^{\infty} \left[e^{i(\omega_0 + C(\xi - \xi_0))t} + e^{i\omega t} - e^{i(\omega_0 + C(\xi - \xi_0))t} \right] e^{-i\xi x} \hat{p}(\xi - \xi_0)\, d\xi. \end{aligned}$$

The integral involving the first term in brackets is just

$$e^{i(\omega_0 t - \xi_0 x)} \int_{-\infty}^{\infty} e^{-i(\xi - \xi_0)(x - Ct)} \hat{p}(\xi - \xi_0)\, d\xi = e^{i(\omega_0 t - \xi_0 x)} p(x - Ct).$$

So we have, using (B.4),

$$\begin{aligned} \left| v^n(x) - e^{i(\omega_0 t - \xi_0 x)} p(x - Ct) \right| &= \left| \int_{-\infty}^{\infty} (e^{i\omega t} - e^{i(\omega_0 + C(\xi - \xi_0))t}) e^{-i\xi x} \hat{p}(\xi - \xi_0)\, d\xi \right| \\ &\le \int_{-\infty}^{\infty} Mt(\xi - \xi_0)^2 |\hat{p}(\xi - \xi_0)|\, d\xi \\ &= Mt \int_{-\infty}^{\infty} |\xi^2 \hat{p}(\xi)|\, d\xi \\ &= Mt \int_{-\infty}^{\infty} |\hat{p''}(\xi)|\, d\xi \;=\; Mt\|\hat{p''}\|_1. \quad \blacksquare \end{aligned}$$

If p is smooth, then the right hand side of (B.6) is small. To make p smooth we will broaden it, while continuing to hold h and k fixed, although the same results could be obtained by leaving p fixed and reducing h and k.

Lemma B.2. *Suppose $p(x) = P(\epsilon x)$ for some fixed function $P \in C_0^2$ with $\hat{P''} \in L_1$. Then*

$$\|\hat{p''}\|_1 = \epsilon^2 \|\hat{P''}\|_1. \tag{B.7}$$

Proof. Define $y = \epsilon x$. Then

$$\begin{aligned} \hat{p''}(\xi) &= \frac{1}{2\pi}\int e^{i\xi x} \frac{d^2 p(x)}{dx^2}\, dx = \frac{1}{2\pi}\int e^{i\xi x} \frac{d^2 P(\epsilon x)}{dx^2}\, dx \\ &= \frac{1}{2\pi}\int e^{i\xi y/\epsilon} \frac{d^2 P(y)}{dy^2} \left(\frac{dy}{dx}\right)^2 dy \left(\frac{dx}{dy}\right) \\ &= \frac{\epsilon}{2\pi}\int e^{i\xi y/\epsilon} P''(y)\, dy = \epsilon \hat{P''}(\xi/\epsilon). \end{aligned}$$

Now define $\eta = \xi/\epsilon$. Then

$$\|\hat{p}''\|_1 = \int |\hat{p}''(\xi)|\,d\xi = \epsilon \int |\hat{P}''(\eta)|\,d\xi$$
$$= \epsilon \int |\hat{P}''(\eta)|\,d\eta \frac{d\xi}{d\eta} = \epsilon^2 \int |\hat{P}''(\eta)|\,d\eta = \epsilon^2 \|\hat{P}''\|_1. \quad \blacksquare$$

Proof of Theorem 4.2.4

Now let $\tilde{Q}$ be a model of an initial boundary problem problem on $x = jh$, $j \geq 0$, consisting of the formula Q described above for $j \geq \ell$ together with the boundary formulas (4.2.3)

$$\sum_{j=0}^{j_{\max}} \sum_{\sigma=-1}^{\sigma_{\max}} S_{j\sigma} v_j^{n-\sigma} = g^n, \tag{B.8}$$

where each $S_{j\sigma}$ is a vector of length ℓ. We assume $v^0 \equiv 0$.

Theorem 4.2.4. *Suppose Q is Cauchy stable but $\tilde{Q}$ admits a strictly rightgoing generalized eigensolution*

$$v_j^n = z^n \sum_{i=1}^{q} a_i \kappa_i^j \tag{B.9}$$

with $|z| = |\kappa_i| = 1$ and $C_i > 0$ for $i = 1, \ldots, q$. Then

$$\|S_{bc}^{(n)}\|_2 \geq \text{const.}\ n \quad \forall n > 0. \tag{B.10}$$

Proof. As described above, the idea of the proof is as follows. We solve the Cauchy problem for Q with initial data $v^0(x)$ whose support is in $x < 0$, obtaining $v^n(x)$ for $n \geq 0$. Then the restriction of v^n to $x = jh$, $j \geq 0$ is identical to the solution that would have been obtained under $\tilde{Q}$ with $v^0 \equiv 0$ and the boundary data $\{g^n\}$ defined by (B.8). In particular, given N, we will pick initial data v^0 such that $\|v^0\|_2 = 0$ and

$$\|v^N\|_2 \geq \text{const.}\ \sqrt{N}, \tag{B.11}$$

where $\|\cdot\|_2$ denotes the discrete ℓ_2 norm (3.6.2) on $j \geq 0$, but such that g satisfies

$$\|g\|_2 \leq \frac{\text{const.}}{\sqrt{N}}. \tag{B.12}$$

These two bounds will then imply (B.10).

Here are the details. Let $P \in C_0^2$ be a fixed function with $P(x) > 0$ on $(-1, 0)$, $P(x) \equiv 0$ elsewhere, and $\hat{P}'' \in L_1$, and write $P'_{\max} = \sup|P'(x)|$. For example P

might be

$$P(x) = \begin{cases} \sin^4 \pi x & x \in [-1, 0], \\ 0 & x \notin [-1, 0]. \end{cases} \tag{B.13}$$

Let N be given and set $T = Nk$. Consider the Cauchy the problem for Q with initial data

$$v^0(x) = \sum_{i=1}^{q} a_i \kappa_i^j p_i(x), \qquad p_i(x) = P(x/C_i T). \tag{B.14}$$

Let M_i be the constant of Lemma B.1 for the wave κ_i, z_i. For any n write $t = nk$. By Lemmas B.1 and B.2, we have then

$$\left| v^n(x) - \sum a_i \kappa_i^j z^n p_i(x - C_i t) \right| \leq t \sum |a_i| M_i \|\hat{p}''\|_1$$
$$= \frac{t \|\hat{P}''\|_1}{T^2} \sum |a_i| M_i C_i^{-2}.$$

In particular, for $n \leq N$ and hence $t \leq T$, this equation together with (B.14) implies

$$\left| v_j^n - \sum a_i \kappa_i^j z^n P\left(\frac{jh}{C_i T} - \frac{nk}{T} \right) \right| \leq \frac{A_1}{T} \tag{B.15}$$

where $A_1 = \|\hat{P}''\|_1 \sum |a_i| M_i C_i^{-2}$.

Now we are equipped to show that $\|g\|_2$ is small, where g is the "equivalent boundary data" (B.8). Given n and $t = nk$, define for all σ and j

$$\tilde{v}_j^{n-\sigma} = P\left(\frac{-t}{T} \right) \sum a_i \kappa_i^j z^{n-\sigma}.$$

Then we have

$$|v_j^{n-\sigma} - \tilde{v}_j^{n-\sigma}| \leq \left| v_j^{n-\sigma} - \sum a_i \kappa_i^j z^{n-\sigma} P\left(\frac{jh}{C_i T} - \frac{t - \sigma k}{T} \right) \right|$$
$$+ \sum |a_i \kappa_i^j z^{n-\sigma}| \left| P\left(\frac{jh}{C_i T} - \frac{t - \sigma k}{T} \right) - P\left(\frac{-t}{T} \right) \right|$$
$$\leq \frac{A_1}{T} + \sum |a_i \kappa_i^j z^{n-\sigma}| P'_{\max} \left| \frac{jh}{C_i T} - \frac{\sigma k}{T} \right|.$$

Therefore for some $A_2 < \infty$,

$$|v_j^{n-\sigma} - \tilde{v}_j^{n-\sigma}| \leq \frac{A_2}{T} \qquad \text{for } 0 \leq j \leq j_{\max}, \quad -1 \leq \sigma \leq \sigma_{\max}. \tag{B.16}$$

Now by definition, $\tilde{v}$ is the generalized eigensolution (B.9) times the constant $P(-t/T)$, which implies

$$\sum_{j=0}^{j_{\max}} \sum_{\sigma=-1}^{\sigma_{\max}} S_{j\sigma} \tilde{v}_j^{n-\sigma} = 0.$$

Consequently we have from (B.8)

$$\|g^n\| \leq \sum_{j=0}^{j_{\max}} \sum_{\sigma=-1}^{\sigma_{\max}} \|S_{j\sigma}\| \, |v_j^{n-\sigma} - \bar{v}_j^{n-\sigma}|. \tag{B.17}$$

By (B.16), each summand on the right is $O(T^{-1})$. Therefore

$$\|g^n\| \leq \frac{A_3}{T} \qquad (n \leq N) \tag{B.18}$$

for some A_3. Hence

$$\|g\|_2^2 = k \sum_{n=1}^{N} \|g^n\|^2 \leq kN\left(\frac{A_3}{T}\right)^2 = \frac{A_3^2}{kN}, \tag{B.19}$$

and taking the square root gives (B.12).

The other half of the argument is to show that $\|v^N\|_2$ is big. Now by definition of the numbers κ_i, we know that the generalized eigensolution (B.9) cannot be zero at more than $\ell - 1$ consecutive grid points without being identically zero. It follows that one has

$$\sum_{j=0}^{\infty} \left| \sum_{i=1}^{q} a_i \kappa_i^j z^n P\left(\frac{jh}{C_i T} - 1\right) \right|^2 > A_4 T \tag{B.20}$$

for some A_4, so long as $T \geq T_0 > \ell h / \max_i C_i$. Squarerooting and using (B.15), we get (B.11), as desired. ∎

Proof of Theorem 4.2.3 (two-level case)

Now we prove

Theorem 4.2.3. *Suppose Q is Cauchy stable but $\bar{Q}$ admits a strictly rightgoing generalized eigensolution (B.9), as before. Then*

$$\|S^n\|_2 \geq \text{const.} \sqrt{n} \tag{B.21}$$

for infinitely many integers $n > 0$.

Proof. The most obvious proof was described in §4.2, especially Fig. 4.3. To adapt that argument to the present framework of considering the Cauchy problem modeled by Q, we could consider the process illustrated in Fig. B.2:

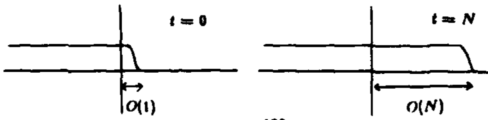


FIG. B.2

It seems clear that this kind of setup should produce growth with respect to initial data proportional to $\sqrt{N}$. However, *no matter how smooth the envelope in Fig. B.2* is, the solution will not satisfy the boundary conditions for $\bar{Q}$ exactly, and we are faced again with the problem of treating "equivalent boundary data." It turns out that this can be done by means of Duhamel's principle, but in the end one gains nothing by having considered the process of Fig. B.2 rather than that of Fig. B.1.

Therefore consider again exactly the setup of the last proof. Let $\{v_j^n\}$ again denote the solution obtained under Q on $(-\infty, \infty)$ with initial data (B.14). Since S is the solution operator for the model $\bar{Q}$ with homogeneous boundary data, we have in general $v^{n+1} \neq S v^n$. However, for each $n \geq 1$, let $\{\tilde{v}_j^{n+1}\}$ be defined by the formula

$$\sum_{j=0}^{j_{\max}} S_{j,-1} \tilde{v}_j^{n+1} = g^n, \tag{B.22}$$

with g^n, as usual, given by (B.8). By Ass. 4.1 (solvability), $\tilde{v}_j^{n+1}$ is a bounded function of g^n, and with (B.19) this implies

$$\|\tilde{v}^{n+1}\| \leq \frac{\text{const.}}{N}. \tag{B.23}$$

Now by (B.8) and (B.22), we have $S v^n = v^{n+1} - \tilde{v}^{n+1}$, that is,

$$v^{n+1} = \tilde{v}^{n+1} + S v^n.$$

Iterating this equation (Duhamel's principle), one obtains

$$v^N = \tilde{v}^N + S\tilde{v}^{N-1} + S^2\tilde{v}^{N-2} + \cdots + S^{N-1}\tilde{v}^1 + S^N v^0, \tag{B.24}$$

where the last term is 0. This implies

$$\|v^N\| \leq N \max_{0 \leq n \leq N-1} \|S^n\| \max_{1 \leq n \leq N} \|\tilde{v}^n\|,$$

hence by (B.23) and (B.11),

$$\max_{0 \leq n \leq N-1} \|S^n\| \geq \text{const.} \|v^N\| \geq \text{const.} \sqrt{N}.$$

This proves (B.21). ∎

Extension to multilevel difference models

To prove Thms. 4.2.3 and 4.2.4 in full generality, we must extend the above arguments to formulas involving vectors rather than scalars and an arbitrary number of levels rather than two. The extension to vectors is straightforward, given Assumption 2.1 (diagonalizability) and the consequent developments of §2.5, §3.6, and §4.2, so we will not discuss it. What we will do is indicate how the extension to multilevel (but scalar) schemes can be treated. We will describe only Lemma B.1, as this is the heart of the proofs.

Let Q be an $s+2$-level scalar difference formula applied on $(-\infty, \infty)$. We can reduce Q to a two-level *model of dimension* $s+1$ in the standard way [Ri67] by introducing the vectors

$$w^n(x) = \left(v^n(x), v^{n+1}(x), \ldots, v^{n+s}(x)\right)^T. \qquad (B.25)$$

If Q has the form (2.1.3), then the equivalent two-level scheme has the structure of a *companion matrix*,

$$w^{n+1}(x) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \mathbf{0} \\ & \mathbf{0} & \ddots & \ddots & \\ & & & 0 & 1 \\ Q_{-1}^{-1}Q_s & \cdots & & Q_{-1}^{-1}Q_1 & Q_{-1}^{-1}Q_0 \end{pmatrix} w^n(x). \qquad (B.26)$$

Taking the Fourier transform, we obtain

$$\hat{w}^{n+1}(\xi) = G(\xi)\hat{w}^n(\xi), \qquad (B.27)$$

where each $\hat{w}^n(\xi)$ is a vector of length $s+1$ and $G(\xi)$ is a square matrix of this size called the *amplification matrix*. By iterating (B.27) and taking the inverse transform, we obtain the representation

$$w^n(x) = \int_{-\infty}^{\infty} G^n(\xi)\hat{w}^0(\xi)e^{-i\xi x}\,d\xi. \qquad (B.28)$$

For any wave number ξ, the eigenvalues of $G(\xi)$ are the associated frequencies ω. Typically there are $s+1$ of these, but for some values of ξ several eigenvalues will come together with multiplicity greater than 1, and $G(\xi)$ will be defective (cf. Thm. 2.1.1). It is this possibility that makes (B.28) *more complicated* than the corresponding scalar formula. However, if ξ_0 and ω_0 are real, then Cauchy stability implies that ω_0 is simple (Thm. 2.2.1), and Thm. 2.3.1 shows further that one can choose $\omega = \omega(\xi)$ with $\omega(\xi_0) = \omega_0$ such that a bound (B.3) is satisfied. These facts make Lemma B.1 extend as follows to multilevel formulas.

Lemma B.1′—multilevel case. *Let $p(x)$ belong to C_0^2 and satisfy $\hat{p}'' \in L_1$. Let Q be applied with initial data*

$$v^n(x) = e^{i(\omega_0 t - \xi_0 x)}p(x - Ct) \qquad n = 0, \ldots, s, \quad t = nk.$$

Then for any $n \geq 0$ and any $x \in \mathbb{R}$, $v^n(x)$ satisfies

$$\left|v^n(x) - e^{i(\omega_0 t - \xi_0 x)}p(x - Ct)\right| \leq \text{const.}\ t\|\hat{p}''\|_1, \qquad (B.29)$$

where $t = nk$.

Proof. The initial data have the vector form

$$w^0(x) = \left(p(x), e^{i\omega_0 k}p(x - Ck), \ldots, e^{i\omega_0 sk}p(x - Csk)\right)^T e^{-i\xi_0 x},$$

and the Fourier transform of this is

$$\hat{w}^0(\xi) = W(\xi)\hat{p}(\xi - \xi_0),$$

where $W(\xi)$ denotes

$$W(\xi) = \left(1, e^{i(\omega_0 + C(\xi - \xi_0))k}, \ldots, e^{i(\omega_0 + C(\xi - \xi_0))sk}\right)^T. \qquad (B.30)$$

By the argument above, $G(\xi)$ has an eigenvalue $e^{i\omega(\xi)k}$ for all ξ such that $\omega(\xi)$ satisfies (B.3) and $\operatorname{Im}\omega(\xi) \geq 0$. *The corresponding eigenvector is*

$$\tilde{W}(\xi) = (1, e^{i\omega(\xi)k} \ldots, e^{i\omega(\xi)sk})^T. \qquad (B.31)$$

Because $\tilde{W}$ is an eigenvector of G, (B.28) can be rewritten

$$\begin{aligned} w^n(x) &= \int_{-\infty}^{\infty} G^n(\xi)\left[\tilde{W}(\xi) + W(\xi) - \tilde{W}(\xi)\right]\hat{p}(\xi - \xi_0)e^{-i\xi x}\,d\xi \\ &= \int_{-\infty}^{\infty}\left[e^{i\omega(\xi)t}\tilde{W}(\xi) + G^n(\xi)(W(\xi) - \tilde{W}(\xi))\right]\hat{p}(\xi - \xi_0)e^{-i\xi x}\,d\xi. \end{aligned}$$

From this expression we need only the first component, which is $v^n(x)$. By (B.31), the first component of the integral of the first term is simply

$$\int_{-\infty}^{\infty} e^{i\omega(\xi)t}\hat{p}(\xi - \xi_0)e^{-i\xi x}\,d\xi.$$

This is exactly the integral we estimated in the proof of Lemma B.1, and we showed that it differs from $e^{i(\omega_0 t - \xi_0 x)}p(x - Ct)$ by at most $Mt\|\hat{p}''\|_1$. Therefore (B.29) will be established if we can bound the integral of the second term correspondingly,

$$\left\| \int_{-\infty}^{\infty} G^n(\xi)(W(\xi) - \tilde{W}(\xi))\hat{p}(\xi - \xi_0)e^{-i\xi x}\, d\xi \right\| \leq \text{const. } t\|\hat{p}''\|_1. \qquad (B.32)$$

Now by Cauchy stability, $\|G^n\|$ is uniformly bounded for all n. Moreover from (B.30), (B.31), (B.3), and the fact used before that $\eta \mapsto e^{i\eta}$ is a contraction map for $\operatorname{Im} \eta \geq 0$, one has

$$\|W(\xi) - \tilde{W}(\xi)\|_\infty \leq \text{const. } (\xi - \xi_0)^2.$$

Eq. (B.32) follows from these facts, since as before we can eliminate the term $(\xi - \xi_0)^2$ by replacing $\hat{p}$ by $\hat{p}''$. ∎

REFERENCES

[Ab79] S. Abarbanel and D. Gottlieb, *Stability of two-dimensional initial boundary value problems using leap-frog type schemes*, Math. Comp. 33 (1979), 1145-55.

[Ab81] S. Abarbanel and E. Murman, *Stability of two-dimensional hyperbolic initial boundary value problems for explicit and implicit schemes*, Proc. NASA-Ames Symp. on Numerical Boundary Procedures, 1981, 199-207.

[Al74] R. Alfold, K. Kelley, and D. Boore, *Accuracy of finite-difference modeling of the acoustic wave equation*, Geophysics 39 (1974), 834-42.

[Ap68] M. Apelkrans, *On difference schemes for hyperbolic equations with discontinuous initial values*, Math. Comp. 22 (1968), 525-39.

[Au73] B. Auld, *Acoustic Fields and Waves in Solids*, Wiley-Interscience, 1973.

[Ba80] A. Bamberger, G. Chavent and P. Lailly, *Etude de schémas numériques pour les équations de l'élastodynamique linéaire*, Res. Rep. 41, INRIA, France, 1980.

[Be79] R. Beam and R. Warming, *An implicit factored scheme for the compressible Navier-Stokes equations II: The numerical ode connection*, Proc. AIAA 4th Comp. Fluid Dynamics Conf., Williamsburg, Va., 1979, 1-13.

[Be81] R. Beam, R. Warming, and H. Yee, *Stability analysis for numerical boundary conditions and implicit difference approximations of hyperbolic equations*, Proc. NASA Ames Symp. on Numerical Boundary Procedures, 1981, 199-207.

[Bo54] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Clarendon Press, Oxford, 1954.

[Br77] P. Brenner, *L_p-estimates of diference schemes for strictly hyperbolic systems with nonsmooth data*, SIAM J. Numer. Anal. 14 (1977), 1126-44.

[Br70] P. Brenner and V. Thomée, *Stability and convergence rates in L_p for certain difference schemes*, Math. Scand. 27 (1970), 5-23.

[Br75] P. Brenner, V. Thomée and L. Wahlbin, *Besov Spaces and Applications to Difference Methods for Initial Value Problems*, Springer Lecture Notes in Mathematics 434, 1975.

[Br81] W. Briggs, T. Saric, and A. Newell, *A new mechanism by which partial* *difference equations can destabilize*, to appear.

[Br53] L. Brillouin, *Wave Propagation in Periodic Structures*, Dover, 1953.

[Br60] L. Brillouin, *Wave Propagation and Group Velocity*, Academic Press, 1960.

[Br79] D. Brown, *Interface approximations for the wave equation*, unpublished, 1979.

[Br73] G. Browning, H.-O. Kreiss and J. Oliger, *Mesh refinement*, Math. Comp. 27 (1973), 29-39.

[Bu78] A. Burns, *A necessary condition for the stability of a difference approximation to a hyperbolic system of partial differential equations*, Math. Comp. 32 (1978), 707-24.

[Ch75] R. Chin, *Dispersion and Gibbs phenomenon associated with difference approximations to initial boundary-value problems for hyperbolic equations*, J. Comp. Phys. 18 (1975), 233-47.

[Ch78] R. Chin and G. Hedstrom, *A dispersion analysis for difference schemes: tables of generalized Airy functions*, Math. Comp. 32 (1978), 1163-70.

[Ch79] R. Chin, G. Hedstrom and K. Karlsson, *A simplified Galerkin method for hyperbolic equations*, Math. Comp. 33 (1979), 647-658.

[Ch83] R. Chin and G. Hedstrom, *A survey of results on modified equations for difference schemes*, J. Comp. Phys., to appear.

[Ch79b] C. K. Chu, *Computational Fluid Dynamics*, in Parter (ed.), Numerical Methods for Partial Differential Equations, Academic Press, 1979, 149-75.

[Ci71] M. Ciment, *Stable difference schemes with uneven mesh spacings*, Math. Comp. 25 (1971), 219-227.

[Ci72] M. Ciment, *Stable matching of difference schemes*, SIAM J. Numer. Anal. 9 (1972), 695-701.

[Cl76] J. Claerbout, *Fundamentals of Geophysical Data Processing*, McGraw-Hill, 1976.

[Cl79] R. Clayton and D. Brown, *The choice of variables for elastic wave extrapolation*, Stanford Exploration Project Rep. No. 20, 1979, 73-96.

[Co80] W. Coughran, Jr., *On the Approximate Solution of Hyperbolic Initial-Boundary Value Problems*, PhD diss., Dept. of Comp. Sci., STAN-CS-80-806, Stanford University, 1980.

[Co62] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, v. II, Wiley-Interscience, 1962.

[En77] B. Engquist and A. Majda, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp. 31 (1977), 629-51.

[Fo73] B. Fornberg, *On the instability of Leap-Frog and Crank-Nicolson approximations of a nonlinear partial differential equation*, Math. Comp. 27 (1973),

45-57.

[Go78] M. Goldberg and E. Tadmor, *Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. I*, Math. Comp. 32 (1978), 1097-1107.

[Go81] M. Goldberg and E. Tadmor, *Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II*, Math. Comp. 36 (1981), 603-626.

[Go78b] D. Gottlieb and E. Turkel, *Boundary conditions for multistep finite-difference methods for time-dependent equations*, J. Comp. Phys. 26 (1978), 181-96.

[Gr81] W. Gropp, *Numerical Solution of Transport Equations*, PhD diss., Dept. of Comp. Sci., Stanford U., 1981.

[Gu75] B. Gustafsson, *The convergence rate for difference approximations to mixed initial boundary value problems*, Math. Comp. 29 (1975), 396-406.

[Gu81] B. Gustafsson, *The choice of numerical boundary conditions for hyperbolic systems*, Proc. NASA Ames Symp. on Numerical Boundary Procedures, 1981, 209-225.

[Gu72] B. Gustafsson, H.-O. Kreiss and A. Sundström, *Stability theory of difference approximations for initial boundary value problems II*, Math. Comp 26 (1972), 649-686.

[Gu82] B. Gustafsson and J. Oliger, *Stable boundary approximations for implicit time discretizations for gas dynamics*, SIAM J. Sci. Stat. Comp., to appear.

[He65] G. Hedstrom, *The near-stability of the Lax-Wendroff method*, Numer. Math. 7 (1965), 73-77.

[He66] G. Hedstrom, *Norms of powers of absolutely convergent Fourier series*, Mich. Math. J. 13 (1966), 393-416.

[He68] G. Hedstrom, *The rate of convergence of some difference schemes*, SIAM J. Numer. Anal. 5 (1968), 363-406.

[He75] G. Hedstrom, *Models of difference schemes for $u_t + u_x = 0$ by partial differential equations*, Math. Comp. 29 (1975), 969-77.

[Je37] F. Jenkins and H. White, *Fundamentals of Physical Optics*, McGraw-Hill, 1937.

[Kr66] H.-O. Kreiss, *Difference approximations for the initial-boundary value problem for hyperbolic differential equations*, in Numerical Solution of Nonlinear Differential Equations, D. Greenspan (ed.), Proc. Adv. Symp. Math. Res. Ctr., U. of Wisconsin, Wiley, 1966.

[Kr68] H.-O. Kreiss, *Stability theory for difference approximations of mixed initial boundary value problems I*, Math. Comp. 22 (1968), 703-714.

[Kr70] H.-O. Kreiss, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math. 23 (1970), 277-298.

[Kr71] H.-O. Kreiss, *Difference approximations for initial boundary-value problems*, Proc. Royal Soc. 329 (1971), 255–61.

[Kr74] H.-O. Kreiss, *Initial boundary value problems for hyperbolic partial differential equations*, Proc. International Congress of Mathematicians, Vancouver, 1974, 127–34.

[Kr74b] H.-O. Kreiss, *Boundary conditions for hyperbolic differential equations*, Proc. Dundee Conf. on Numer. Soln. of Diffl. Eqs., Springer-Verlag Lect. Notes in Math. 363, 1974.

[Kr68b] H.-O. Kreiss and E. Lundqvist, *On difference approximations with wrong boundary values*, Math. Comp. 22 (1968), 1–12.

[Kr72] H.-O. Kreiss and J. Oliger, *Comparison of accurate methods for the integration of hyperbolic equations*, Tellus 24 (1972), 199–215.

[Kr73] H.-O. Kreiss and J. Oliger, *Methods for the Approximate Solutions of Time Dependent Problems*, Global Atmospheric Research Programme Publ. Ser. no. 10, Geneva, 1973.

[Li78] J. Lighthill, *Waves in Fluids*, Cambridge University Press, 1978.

[Lu81] B. Lucier, *Dispersive Approximations for Hyperbolic Conservation Laws*, PhD diss., Dept. of Math., U. of Chicago, 1981.

[Ma69] A. Maradudin, et al., *Lattice Dynamics*, W. A. Benjamin, 1969.

[Ma81] L. Martineau-Nicoletis, *Simulations numérique de la propagation d'ondes sismiques dans les milieux stratifi'es á deux et trois dimensions*, PhD diss., Univ. Pierre and Marie Curie, Paris, 1981.

[Mi80] D. Michelson, *Initial boundary-value problems for hyperbolic equations and their difference approximation with characteristic boundary*, PhD diss., Dept. of Math. Sci., Tel-Aviv U., 1980.

[Mi81] D. Michelson, *Stability theory of difference approximations for multidimensional initial-boundary value problems*, to appear.

[Mo53] P. Morse and H. Feshbach, *Methods of Theoretical Physics*, v. I, McGraw-Hill, 1953.

[Ok76] M. Okrouhlik and R. Brepta, *Side effects of finite element method applied on stress wave propagation in a thin elastic bar*, Acta. Technica Čsav. 4 (1976), 417–438.

[Ol74] J. Oliger, *Fourth-order difference methods for the initial boundary-value problem for hyperbolic equations*, Math. Comp. 28 (1974), 15–25.

[Ol76] J. Oliger, *Hybrid difference methods for the initial boundary-value problem for hyperbolic equations*, Math. Comp. 30 (1976), 724–38.

[Ol79] J. Oliger, *Constructing stable difference methods for hyperbolic equations*, in S. Parter (ed.), Numerical Methods for Partial Differential Equations, Academic Press, 1979, 255–71.

[Or72] J. Ortega, *Numerical Analysis: A Second Course*, Academic Press, 1972.

[Os69a] S. Osher, *Stability of difference approximations of dissipative type for mixed initial-boundary value problems. I*, Math. Comp. 23 (1969), 567–72.

[Os69b] S. Osher, *Systems of difference equations with general homogeneous boundary conditions*, Trans. Amer. Math. Soc. 137 (1969), 177–201.

[Os69c] S. Osher, *On systems of difference equations with wrong boundary conditions*, Math. Comp. 23 (1969), 567–72.

[Os72] S. Osher, *Stability of parabolic difference approximations to certain mixed initial boundary value problems*, Math. Comp. 26 (1972), 13–39.

[Pe76] J. Peetre, *New Thoughts on Besov Spaces*, Duke U. Math. Series I, 1976.

[Ra72] J. Rauch, *L_2 is a continuable initial condition for Kreiss' mixed problems*, Comm. Pure Appl. Math. 25 (1972), 265–95.

[Ri67] R. Richtmyer and K. Morton, *Difference Methods for Initial-value Problems*, Wiley-Interscience, 1967.

[Se63] S. Serdjukova, *A study of stability in C of explicit difference schemes with constant real coefficients which are stable in ℓ_2*, Ž. Vychisl. Mat. Mat. Fiz. 3 (1963), 365–370.

[Se66] S. Serdjukova, *On the stability in C of linear difference schemes with constant real coefficients*, Ž. Vychisl. Mat. Mat. Fiz. 6 (1966), 477–486.

[Sk79] G. Sköllermo, *Error analysis of finite difference schemes applied to hyperbolic initial boundary value problems*, Math. Comp. 33 (1979), 11–35.

[So64] A. Sommerfeld, *Optics*, Academic Press, 1964.

[St63] H. Stetter, *Maximum bounds for the solutions of initial value problems for partial difference equations*, Numer. Math. 5 (1963), 399–424.

[St62] G. Strang, *Trigonometric polynomials and difference methods of maximum accuracy*, J. Math. Phys. 41 (1962), 147–154.

[St64] G. Strang, *Wiener-Hopf difference equations*, J. Math. Mech. 13 (1964), 85–96.

[St66] G. Strang, *Implicit difference methods for initial-boundary value problems*, J. Math. Anal. Applic. 16 (1966), 188–198.

[Su74] A. Sundström, *Efficient numerical methods for solving wave propagation equations for non-homogeneous media*, Swedish Defense Institute Report FOA 4 C 4576-A2, 1974.

[Ta78] E. Tadmor, *Scheme-independent approximations to hyperbolic initial-boundary value systems*, PhD diss., Dept. of Math. Sci., Tel-Aviv U., 1978.

[Ta81] E. Tadmor, *Unconditional instability of inflow-dependent boundary conditions in difference approximations to hyperbolic systems*, Proc. NASA Ames Symp. Numerical Boundary Procedures, 1981, 323–332.

[Th65] V. Thomée, *Stability of difference schemes in the maximum norm,* J. Diffl. Eqs. 1 (1965), 272-292.

[Th69] V. Thomée, *Stability theory for partial difference operators,* SIAM Review 11 (1969), 152-195.

[Tr82] L. Trefethen, *Group velocity for finite difference schemes,* SIAM Review, to appear.

[Tr83] L. Trefethen, *Group velocity interpretation of the stability theory of Gustafsson, Kreiss, and Sundström,* J. Comp. Phys., to appear.

[Va70] J. Varah, *Maximum norm stability of difference approximations to the mixed initial boundary-value problem for the heat equation,* Math. Comp. 24 (1970), 31-44.

[Va71] J. Varah, *Stability of difference approximations to the mixed initial boundary value problem for parabolic systems,* SIAM J. Numer. Anal. 8 (1971), 598-615.

[Vi75] R. Vichnevetsky and B. Peiffer, *Error waves in finite element and finite difference methods for hyperbolic equations,* in Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky (ed.), Assoc. Int. Calcul Analogique, Ghent, Belgium, 1975, 53-58.

[Vi80] R. Vichnevetsky, *Propagation characteristics of semi-discretizations of hyperbolic equations,* Math. and Computers in Simulation 22 (1980), 98-107.

[Vi81] R. Vichnevetsky, *Energy and group velocity in semi discretizations of hyperbolic equations,* Math. and Comp. in Simulation 23 (1981), 333-43.

[Vi81b] R. Vichnevetsky, *Propagation through numerical mesh refinement for hyperbolic equations,* Math. and Comp. in Simulation 23 (1981), 344-53.

[Vi82] R. Vichnevetsky and J. Bowles, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations,* SIAM, to appear, 1982.

[Wa74] R. Warming and B. Hyett, *The modified equation approach to the stability and accuracy analysis of finite-difference models,* J. Comp. Phys. 14 (1974), 159-79.

[Wa80] J. Watts and W. Silliman, *Numerical dispersion and the origins of the grid-orientation effect: a summery,* Proc. 73rd Annual Meeting, AIChE, Chicago, 1980.

[Wh61] G. Whitham, *Group velocity and energy propagation for three-dimensional waves,* Comm. Pure Appl. Math. 14 (1961), 675-91.

[Wh74] G. Whitham, *Linear and Nonlinear Waves,* Wiley-Interscience, 1974.

INDEX

MED

8